PB86-114105

Ambient Ozone and Human Health
An Epidemiological Analysis. Volume 3

Resources for the Future, Inc., Washington, DC

Prepared for

Environmental Protection Agency
Research Triangle Park, NC

Jun 85

U.S. Department of Commerce
National Technical Information Service
NTIS

PB86-114105

United States
Environmental Protection
Agency

Office of Air Quality
Planning and Standards
Research Triangle Park NC 27711

EPA-450/5-85-005c
August 1985

Air

EPA

Ambient Ozone And Human Health: An Epidemiological Analysis

Volume III

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA. 22161

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before completing)

1. REPORT NO.
EPA-450/5-85-005c

2.

3. RECIPIENT'S ACCESSION NO.
PB86 1141057AS

4. TITLE AND SUBTITLE
Ambient Ozone and Human Health: An Epidemiological Analysis Volume III

5. REPORT DATE
June 1985 (Date of Preparation)

6. PERFORMING ORGANIZATION CODE

7. AUTHOR(S)
Paul R. Portney and John Mullahy

8. PERFORMING ORGANIZATION REPORT NO.

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Resources for the Future
1616 P Street N.W.
Washington, DC 20036

10. PROGRAM ELEMENT NO.
12A2A

11. CONTRACT/GRANT NO.
68-02-3583

12. SPONSORING AGENCY NAME AND ADDRESS
U.S. Enviornmental Protection Agency
Office of Air Quality Planning and Standards (MD-12)
Research Triangle Park, NC 27711

13. TYPE OF REPORT AND PERIOD COVERED
Final Report

14. SPONSORING AGENCY CODE
OAQPS

15. SUPPLEMENTARY NOTES
Project Officer: Thomas G. Walton

16. ABSTRACT

This report is the third volume of an analysis of the relationship between ozone and human health benefits.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group
Benefit Analysis Air Pollution, O_3 Epidemiology		

18. DISTRIBUTION STATEMENT
Release Unlimited

19. SECURITY CLASS *(This Report)*
Unclassified

20. SECURITY CLASS *(This page)*
Unclassified

21. NO. OF PAGES

22. PRICE

EPA Form 2220-1 (Rev. 4-77) PREVIOUS EDITION IS OBSOLETE

AMBIENT OZONE AND HUMAN HEALTH:
AN EPIDEMIOLOGICAL ANALYSIS

Paul R. Portney and John Mullahy

Resources for the Future
1616 P Street, N.W.
Washington, D.C. 20036

Volume III
Final Report
June 1985

Submitted to the Economic Analysis Branch, Office of Air Quality Planning and Standards, Environmental Protection Agency, Research Triangle Park, North Carolina 27711, under contract number 68-02-3583.

2<

DISCLAIMER

This report has been reviewed by the Office of Air Quality Planning and Standards, U. S. Environmental Protection Agency, and approved for publication as received from Resources for the Future. The analysis and conclusions presented in this report are those of the authors and should not be interpreted as necessarily reflecting the official policies of the U. S. Environmental Protection Agency.

TABLE OF CONTENTS

Page

CHAPTER 1. INTRODUCTION 1-1

CHAPTER 2. ECONOMETRIC ESTIMATION OF HEALTH STATUS MODELS

2.1 Introduction 2-1
2.2 Some Problems with Least-Squares Estimation of Health Status Models 2-3
2.3 Tobit Health Outcome Models 2-7
2.4 Cragg-class Health Outcome Models 2-11
2.5 Truncated-Normal Estimation 2-15
2.6 Heckman's Approach: Sample Selection 2-19
2.7 Tobin, Cragg and Heckman: A Digression 2-21
2.8 Poisson-distributed Health Outcome Measures 2-30
2.9 Geometric-distributed Health Outcome Measures 2-33
2.10 Multinomial-distributed Health Outcome Measures 2-35
2.11 Estimation of Grouped Data Models Under the Normality Assumption 2-38
2.12 Summary and Conclusions 2-40

CHAPTER 3. AIR POLLUTION MONITORS AND INDIVIDUAL EXPOSURE 3-1

CHAPTER 4. URBAN AIR QUALITY AND ACUTE RESPIRATORY ILLNESS

4.1 Introduction 4-1
4.2 Framework for the Analysis 4-3
4.3 Model Specification 4-8
4.4 Empirical Results 4-11
4.5 Policy Implications 4-20
Appendix 4-28

CHAPTER 5. CONSTRUCTING A LIFETIME SMOKING PROFILE USING THE 1979 HEALTH INTERVIEW SURVEY 5-1

CHAPTER 6. CIGARETTE SMOKING, AIR POLLUTION, AND RESPIRATORY ILLNESS: AN ANALYSIS

6.1 Introduction 6-1
6.2 Smoking, Pollution, and Acute Illness 6-2
6.3 Data and Estimation Strategy 6-5
6.4 Estimates of Model Parameters and Relative Risks 6-14
Appendix 6-23

CHAPTER 7. CHRONIC RESPIRATORY DISEASE 7-1

4<

Page

CHAPTER 8. ADDITIONAL SENSITIVITY ANALYSES

8.1 The Effects of Precipitation on Acute Health Status 8-1
8.2 Sample Size, Model Specification, and Parameter Estimate Sensitivity 8-2
8.3 Poisson Regression Analysis of Volume I Models (48), (49), and (50) 8-4
8.4 Sensitivity to Aggregation Across Smoking and Chronic Illness Status 8-6

Chapter 1

INTRODUCTION

Volume I of this report presents a great many results from our basic analysis of ozone and acute and chronic illness. As indicated in Volume I, we had to make a number of decisions along the way in the early stages of our research. One of the most important concerned the tradeoff between the breadth of our analysis as opposed to the possible in-depth investigation of a relatively small number of hypotheses. In other words, should we use fairly standard statistical techniques to investigate dose-response relationships for a broad range of possible illness, using a variety of explanatory variables, and in a number of different population groups? Or should we winnow out a relatively few "promising" relationships using preliminary tests, and then allocate time and computing resources to the application of more powerful statistical techniques to these relationships?

With some exceptions we adopted the former approach. Because of our unique and comprehensive data on air pollution concentrations and individuals' health and socioeconomic status, and we elected to test a wide variety of hypotheses about the possible relationships between ozone and other air pollutants on the one hand, and a variety of acute and chronic illnesses on the other. In addition, we examined separately several different degrees of severity for the acute illnesses we examined and we also conducted separate analyses for adults and for children (aged 17 and below). Of course, as we point out in Volume I, we did conduct additional sensitivity analyses where our preliminary research suggested statistically significant associations between ozone and the dependent variable in question. Nevertheless, the general approach was a "broad brush" one.

Since completing the work reported in Volumes I and II, we have received many helpful comments on and constructive criticisms of the approach we took in our analyses. Many of these comments came in an EPA-sponsored public Peer Review Meeting held in Raleigh (N.C.) on April 3, 1984. There experts from the epidemiological, clinical, biostatistical, and economic communities presented us with a number of useful suggestions for further work. In addition, we have received many useful comments from our EPA project officers and from our colleagues at RFF and elsewhere who have read with interest our original work. Finally, we have given considerable thought ourselves to ways in which the original analysis might be extended or improved.

Thus, over the past year we have tried to conduct additional analyses that address some of most important questions arising out of our original work. Volume III below presents the results of some of that work. We say "some" because we are continuing to conduct additional epidemiological analyses as time and resources permit, at least some of which may not be complete until after this report has been submitted. In one sense, in fact, we hope to never be "done" with our work even though this report completes our analysis for EPA.

In one way or another, each of the following chapters is designed to address one or more of the questions raised in our earlier work. For instance, Chapter 2 is purely methodological. It presents a variety of different estimation techniques that may be appropriate when the assumptions that lie behind ordinary least squares (OLS) are violated as several careful readers of the studies in Volume I suggested they might be. There we consider the sorts of problems that arise with OLS in the special context of health effects estimation. Among the alternatives to OLS we consider are Tobit estimation, Cragg-type "hurdles" models, sample-selection and count-data models, multinomial logit approaches, and grouped dependent variable techniques. This chapter is a long and technical one, we realize. However, we feel it is necessary to set the stage for the empirical work presented in later chapters; it should also prove useful to anyone about to embark for the first time on his own estimation of air pollution (or other environmentally-related) health effects.

Chapter 3 is much shorter and simpler. It addresses a common reaction to our original analysis. Remember that in Volume I, the air pollution readings we assign to each individual are those measured at the monitor nearest his or her home, provided that the monitor in question is no more than twenty miles away (sometimes less). We continue to believe this is preferable to the most common alternative to this approach--suggested in the literature matching each individual in an SMSA to the air pollution concentrations averaged over all the monitors in the SMSA, or within a subset of it. However, because most individuals do travel about within an

8<

area, it is possible that the area wide averaged approach might better characterize the exposures of at least some individuals. If so, these averaged concentrations would be the appropriate ones to use in epidemiological analyses. Hence, it is of interest to know how closely correlated are the readings at the monitor(s) nearest the individuals' dwellings with the average of all the monitors within a given radius of the dwelling. This analysis is undertaken in Chapter 3.

That exercise in turn forms the basis for some sensitivity analysis we conduct in Chapter 4 of the effect on our findings of different rules about matching air pollution concentrations to individuals. Chapter 4 extends and improves upon our original work in a number of other ways, as well. For instance, building on the methodology presented in Chapter 2 of this volume, in Chapter 4 we investigate the determinants of acute respiratory disease using poisson regression instead of the OLS and logit techniques employed in Volume I. For reasons presented in Chapters 2 and 4, we consider this to be a significant improvement on our earlier analysis. In addition, Chapter 4 presents a more sophisticated analysis of the possible non-linearities that may characterize the dose-response relationship linking acute respiratory disease to ambient ozone and sulfate concentrations. Not only do we consider spline-type functional relationships, but we also allow for a variety of non-linearities within the (already non-linear) poisson approach. This, too, sheds additional light on the analysis in Volume I. Finally, we believe that the elasticity-of-response calculations contained in the last part of Chapter 4 are a useful way to view the possible effects of changes in ambient ozone concentrations on human health. This suggests how our findings might be used in applied policy analysis if it were desired to do so.

One of the respects in which the analysis in Volume I could clearly be improved concerns the measures of cigarette smoking we employed. Recall that in most of the models estimated, we used NCIGS, a continuous measure of daily cigarette consumption, or SMOKY1NO, a dummy variable indicating whether or not an individual is a never- or former smoker as opposed to a current smoker. We also occasionally used an additional dummy variable, FORMER, to distinguish between those who do not smoke now but once did from those who never smoked.

However, even this additional treatment resulted in our finding a less pronounced relationship between smoking and ill health than we might have expected (although we hasten to point out that even our crude measures of smoking were often positively and significantly associated with ill health). One reason for this was our inability in the Volume I analyses to make use of all the data provided in the HIS Smoking Supplement on individuals' lifetime smoking histories. Thus, one of our purposes in the analyses we have conducted since April 1984 was to develop a measure of lifetime smoking behavior and employ it in our analyses. Chapter 5 presents the approach we took in doing so. While the HIS smoking data do not enable us to specify an exact profile of respondents' lifetime smoking habits, they do permit the construction of several plausible profiles. These are discussed in some detail in Chapter 5. Among other things, that chapter discusses the differing weights that might be given to cigarettes smoked years ago compared to recent cigarette consumption.

Chapter 6 presents the results of additional empirical analysis of the relationship between air pollution (ozone and sulfates) and acute respiratory disease. It extends the analysis in Chapter 4 of this volume, and all the work in Volume I, in several important ways. First, the analysis in Chapter 6 incorporates the more sophisticated measures of individual smoking. For instance, in addition to NCIGS, a measure of current smoking habit, the analysis also includes the variable PACKS, a proxy for lifetime cigarette consumption. The analysis in Chapter 6 extends our earlier research in another suggested direction. That is, it models the individuals' health outcomes as a multinomial logit process in which, on any given day during the two-week recall period, an individual could report no restriction of activity at all, a minor restriction in activity attributable to respiratory illness (with no bed confinement), or what we refer to as a "severe" respiratory restriction--i.e., one which requires confinement to bed for at least half the day. For reasons spelled out in Chapter 6, we feel this is another productive way to model the possible relationship between ambient air quality and acute respiratory disease. (The chapter also contains a very brief discussion of ordered logit as an estimation approach.)

Chapter 6 is intended to accomplish one additional objective. The comments on our work in Volume I often expressed surprise that cigarette smoking did not completely "swamp" ambient ozone pollution in its contribution to acute (and chronic) illness--even though we generally found a positive and significant association between smoking and illness. Thus, one purpose of the analysis in Chapter 6 is to explore in somewhat greater detail the relative risks posed by cigarette smoking and air pollution. This is important since considerable public resources are currently devoted to reducing both. While far from being comprehensive on the subject, Chapter 6 does explore these relative risks in some detail.

As the preceding pages suggest, most of the emphasis in Volume III falls on the possible relationship between ambient ozone (and sulfate) concentrations and acute respiratory health. This reflects the heavy emphasis given acute health effects in our earlier work as described in Volume I. However, we did devote some attention in Volume I to possible relationships between long-term exposures to air pollution and the prevalence of chronic respiratory and other kinds of disease.

Chapter 7 below presents the results of some preliminary reanalysis of those finding, specifically those dealing with chronic respiratory disease. The analysis below extends our original work in several important ways. First, we restrict our attention in Chapter 7 to a group of individuals who at the time of the 1979 HIS had lived in their present location for at least ten years. This is a more "residentially stable" group than that analyzed in Volume I, an important consideration in the epidemiological investigation of chronic illness. In addition, the individuals analyzed in Chapter 7 are divided into two distinct groups depending on whether or not they received a special supplement (or "probe") on respiratory disease as part of the 1979 HIS. Because the reported incidence of chronic respiratory disease varies by a factor of six between those who received the probe and those who did not, we felt the two groups should be analyzed separately rather than pooled as in our original analysis. Finally, this reanalysis includes some model specifications in which ozone is measured by the ambient concentration averaged over all the monitors with ten or twenty miles of each resident's dwelling.

Finally, in Chapter 8 we report our responses to a variety of comments or queries on Volumes I and II. None of these required the preparation of a separate chapter, but each was important enough to merit consideration.

One final note about Volume III. Several of the chapters have been written to serve more than one purpose. For instance, a slightly revised version of Chapter 4 will be appearing in the Journal of Urban Economics in 1986 under the title "Urban Air Quality and Acute Respiratory Illness." Similarly, the material in Chapter 6 formed the basis of a paper presented at the 1984 annual meetings of the American Economic Association in Dallas. While we have modified them for incorporation into Volume III, some material--particularly the brief descriptions of the HIS and air pollution data bases--will occasionally appear repetitive.

Chapter 2

ECONOMETRIC ESTIMATION OF HEALTH STATUS MODELS

2.1 INTRODUCTION

In the last decade or so, estimation of microeconomic models of individual behavior using large individual- or household-level data sets has flourished and proven an important advance in applied economics. Details typically masked in aggregate time-series data analysis are often available in individual cross-sectional data, thus enabling the testing of hypotheses about responses of individuals to changes in constraints.

In such micro datasets one is prone to find measures that economists would characterize either as corner-solution realizations of instantaneous optimizing decisions or as discrete representations of such decisions. An example of the former case would be where one has data on the number of hours an individual worked in the market over a given year, and for some subset of individuals no market hours were worked.

An instance of the latter case is where data are available only on whether or not an individual had purchased some consumer durable over the previous twelve months, but not on the amount of the expenditure. Assuming such statistical models to be the objectives of estimation, then the former is an example of what have come to be known as limited dependent variable (LDV) models, while the latter is a member of the class of qualitative dependent variable (QDV) models. Tobin's pioneering 1957 paper on durables demand is the forerunner of LDV estimation in economics. Using data on 735 households, Tobin modeled the ratio of durables expenditures to disposable income; for 183 of these spending units, no durables were purchased during the time period of interest and a "corner solution" had to be treated. As is well known, the solution to this problem was the genesis of the Tobit estimator, which will be discussed below. Note that if Tobin only had data on whether or not there was some durable purchased rather than on the actual amount, a QDV model (such as binary probit or logit) would have been the appropriate approach.

In this chapter we discuss the theory and practice of econometric estimation of LDV and QDV models as they pertain to health status measures such as respiratory-related restricted activity days, or the presence-absence of a chronic respiratory condition. It is seen that, owing to the nature of the available micro data, standard econometric techniques such as ordinary least squares (OLS) will typically be inappropriate tools for the analysis of the relationships between air pollution and human health. The available data on health status measures, rather, are generally of a nature best described as qualitative or limited dependent variables. This being the case, more complicated estimation techniques are in general required in order to obtain consistent estimates of the parameters governing the health status outcomes. Maximum likelihood is the estimation method most commonly used in such analysis.

The treatment here is necessarily brief. However, several excellent surveys are available for the reader who wishes more detailed treatments of the topics to be discussed below. The 1981 and 1984 surveys by Amemiya are excellent overviews of qualitative and limited dependent variable models, respectively, and the 1983 monograph by Maddala provides broad coverage in both these areas. The often-cited 1981 volume edited by Manski and McFadden is also an excellent survey of topics in qualitative and limited dependent variable estimation.

15<

Some definitional preliminaries are appropriate here. First, standard practice is followed, with random variables represented in upper-case notation, their realizations in lower-case. Second, the terms "censored distribution" and "truncated distribution" will be used with considerable frequency below. The introduction to chapter 6 of Maddala (1983) provides a good heuristic explanation of censoring and truncation as they pertain to the normal econometric model.

The plan for the remainder of this chapter is as follows. First, we briefly assess problems associated with least squares estimation of air pollution - health status models. Then we turn to a discussion of some techniques that might be considered more or less appropriate for the estimation problems attendant to estimation of health status models. Following this we turn to a discussion of prediction based on the estimation of the various models. A summary concludes the chapter.

2.2 SOME PROBLEMS WITH LEAST-SQUARES ESTIMATION OF HEALTH STATUS MODELS

As mentioned above, this chapter surveys various econometric techniques for estimating health outcome models. As will be seen throughout, these techniques are generally such that iterative (and sometimes costly) maximum likelihood methods are required in order to obtain consistent and efficient estimates of the models' parameters. Since sound econometric policy analysis depends at least in part on obtaining consistent, if not efficient, parameter estimates, the question is then begged: why is it necessary to utilize such complicated and expensive methods when simple and inexpensive least-squares algorithms abound? In a nutshell, the answer is that least-squares estimates of models of the genre we are considering will generally be biased and inconsistent. The purpose of this section is to briefly demonstrate why this is so. To this end, a brief exposition of the fundamentals of the basic linear econometric model is presented, the requirements for consistent estimation of the parameters are explained, and why at least some of these requirements are unlikely to be met in the health status models to be considered is discussed. The exposition of the linear model and its properties follows that of Schmidt (1976), which is among the most lucid in published texts.

Of fundamental concern is consistent and, if possible, efficient estimation of the parameter vector β in the case where random variables Y_i are distributed in some manner with finite mean μ_i and finite variance σ^2. Specifying $\mu_i = X_i\beta$ makes the problem nontrivial, with X_i a 1xk vector of independent variables which will in general include measures of air pollution and other covariates, and β a kx1 vector of unknown parameters to be estimated. Given these assumptions, we can write

$$Y_i = X_i\beta + \varepsilon_i, \tag{1}$$

where, because $E(Y_i) = X_i\beta$ and $Var(Y_i) = \sigma^2$, ε_i has mean zero and variance σ^2. The unobserved realizations of ε_i correspond to the observed realizations of Y_i, y_i. It is assumed that there exist T independent observations on (y_i, X_i).

The model described satisfies full ideal conditions (Schmidt, p. 2) when

i) X is a nonstochastic matrix of rank $k<T$, and has the property that $Q = \lim_{T\to\infty} X'X/T$ is finite and nonsingular;

ii) $\varepsilon = [\varepsilon_i]$ is distributed multivariate $N(0, \sigma^2 I_T)$. Here X is the Txk matrix of independent variables. y will henceforth denote the Tx1 vector of the realizations y_i.

It can be demonstrated that, with or without the assumption of normality for ε, the OLS estimator of β, $\hat{\beta} = (X'X)^{-1}X'y$, is consistent:

i) $\hat{\beta} = (X'X)^{-1}x'y$

$= (X'X)^{-1}X'(X\beta+\varepsilon)$

$= \beta + (X'X)^{-1}X'\varepsilon$

$E(\hat{\beta}) = \beta + (X'X)^{-1}X'E(\varepsilon)$

$= \beta+0$

$= \beta$, so that $\hat{\beta}$ is unbiased for β;

ii) The covariance matrix of $\hat{\beta}$ is $\sigma^2(X'X)^{-1}$ so that, with all limits taken for $T\to\infty$,

$$\begin{aligned} \lim \sigma^2(X'X)^{-1} &= \sigma^2\lim(X'X)^{-1} \\ &= \sigma^2\lim(X'X/T)^{-1}T^{-1} \\ &= \sigma^2\lim(Q^{-1}T^{-1}) \\ &= 0, \end{aligned}$$

because from above Q is finite nonsingular so that its inverse exists and is finite, and σ^2 is finite;

iii) Therefore, since $\hat{\beta}$ is unbiased and its covariance matrix vanishes in the limit, then $\hat{\beta}$ is consistent.

Because of its computational ease, least squares is obviously an appealing tool for model estimation. The analyst must assess whether any or all of the above conditions fail to characterize the data or model under consideration to see if least squares maintains its consistency properties. Should least squares prove inconsistent, alternative, and generally more costly, methods of estimation must be utilized in order to obtain consistent estimates of β.

As discussed in detail below, a very general characterization of quantitative health outcomes measures is that they are data bounded from below by zero, i.e. data realized only in nonnegative quantities. Of specific concern here are measures like "amount of time spent ill." Such measures are generally modeled econometrically as the censored or truncated counterparts of normally-distributed latent random variables Y_i^* having $E(Y_i) = X_i\beta$, $Var(Y_i) = \sigma^2$. However, if the realizations of Y_i are censored from below at zero, we have

$$E(Y_i^*|y_i>0) = X_i\beta + \sigma\phi_i/\Phi_i, \qquad (2)$$

$$E(Y_i^*) = X_i\beta\Phi_i + \sigma\phi_i,$$

where ϕ_i and Φ_i are the standard normal density and distribution functions evaluated at $(X_i\beta/\sigma)$. In the truncated case, where $Pr(y_i>0) = 1$,

$$E(Y_i^*) = X_i\beta + \sigma\phi_i/\Phi_i. \qquad (3)$$

When defined in terms of these expectations, the problems inherent in least squares estimation become apparent. Since $E(\sigma\phi_i/\Phi_i) \neq 0$, then $E(\varepsilon_i) \neq 0$ when ε_i is defined as the difference between either $E(Y_i)$ or $E(Y_i|y_i>0)$ and $X_i\beta$ in (2). Thus least squares regression of y on X will yield inconsistent estimates of β, given that the null error expectation assumption has been violated. Heckman (1976) is a good general discussion of such problems.

Not all measures of interest in our analysis are cast in terms of normally-distributed, partially-observed random variables. In the other cases we shall investigate, there are yet different characteristics of the data or the assumed statistical distributions that render least squares inappropriate given the objective of consistent parameter estimation. For example, least-squares estimation strategy is generally completely inappropriate when outcomes are qualitative since no objective function of interest can be cast in terms of linear expectations functions like those above. We now turn to an assessment of various approaches to the estimation of health status models.

2.3 TOBIT HEALTH OUTCOME MODELS

A logical starting point is the basic Tobit model. The nature of several of the health status measures of interest in the micro data sets being analyzed in this study is such that Tobit estimation would seem—at least at first blush—to be a sensible approach. (See Ostro (1983) for an application of Tobit to a similar problem.)

Tobit estimation has been utilized in a variety of areas in applied microeconomics, ranging from labor supply (see the excellent survey by Killingsworth (1983)), to health economics (Ostro, (1983)), to commodity demands or expenditures (Tobin (1957), Pitt (1983)), and many others (see Amemiya (1984) for an extensive bibliography). The basic idea underlying Tobit estimation is that one posits the existence of (latent) random variables Y_i^* are independently, normally distributed (NID) $(X_i\beta,\sigma^2)$. In many interpretations of the Tobit model, the Y_i^*, are stochastic indicators of intensity of desire for undertaking some activity. Owing to the nature of the activity, however, some realizations of the Y_i^* are censored while for the others, the intensities are mapped directly into actual undertakings of the activity. Some threshold, in effect, is crossed such that the activities are actually undertaken. For example the fundamental idea behind Tobin's seminal paper is that the Y_i^* represent intensities of desire to purchase consumer durables. When certain (assumed known) thresholds are crossed, these intensities become actual purchases. In most applied areas, the thresholds are zero, so that the mappings from intensities into undertaken activities can be looked at as occurring when the realizations of the Y_i^* occur in the interior of commodity space. Otherwise, corner solutions obtain (for one discussion of estimation in the Kuhn-Tucker/corner-solution/Tobit context, see Wales and Woodland (1983)).

Assuming, then, that the thresholds are known and constant across individuals, the basic Tobit model can be described by (4):

$$\begin{aligned} Y_i^* &\sim \mathrm{NID}(X_i\beta, \sigma^2) \\ y_i &= \max(C, y_i^*). \end{aligned} \tag{4}$$

Setting C = 0 gives the model we shall discuss in the sequel. Letting Ω_0 signify the index set for observations for which $\max(0, y_i^*) = 0$, and Ω_1 be the index set for observations for which $\max(0, y_i^*) > 0$, then the likelihood function for the Tobit model described here is

$$L = \prod_{i\varepsilon\Omega_0}(1-\Phi(\frac{X_i\beta}{\sigma})) \prod_{i\varepsilon\Omega_1}(\phi(\frac{y_i-X_i\beta}{\sigma})/\sigma), \tag{5}$$

In log form (5) is

$$\ell = \sum_{i\varepsilon\Omega_0} \ln(1-\Phi_i) - |\Omega_1|\ln\sigma - \sum_{i\varepsilon\Omega_1}\ell n\phi_i \tag{6}$$

where $|\cdot|$ denotes cardinality and where terms not involving (β,σ) are dropped.

The first-order conditions for maximizing ℓ are the (k + 1) equations

$$\partial\ell/\partial\beta = \sum_{i\varepsilon\Omega_0} (-\lambda_i/\sigma)X_i' + \sum_{i\varepsilon\Omega 1} (y_i-X_i\beta)X_i'/\sigma^2 = 0$$

$$\partial \ell / \partial \sigma = \sum_{i \in \Omega_0} \lambda_i (X_i \beta)/\sigma^2 + \sum_{i \in \Omega_1} ((y_i - X_i \beta)^2 - \sigma^2)/\sigma^3 = 0, \tag{7}$$

where $\lambda_i = \phi_i/(1-\Phi_i)$. Using terms in these equations, the method of Berndt-Hall-Hall-Hausman (1974) among others, can be used for optimization, and statistical inference is based on the asymptotic t-tests generated by utilizing $[\sum_{i=1}^{N}(\ell_i \ell_i')]^{-1}$ as the estimate of $\mathrm{cov}(\hat{\beta})$ (ℓ_i is the i-th term of $[(\partial \ell/\partial \beta)', [\partial \ell/\partial \sigma)]')$.

Several characteristics of the Tobit model are noteworthy. First, as Amemiya (1984) points out, the likelihood function (5) can be rewritten as

$$L = [\prod_{i \in \Omega_0}(1-\Phi_i) \prod_{i \in \Omega_1} \Phi_i] [\prod_{i \in \Omega_1}(\phi_i/\Phi_i \sigma)] \tag{8}$$

Written in this form, the likelihood function of the Tobit model can be viewed as the product of the likelihood functions of a probit model with parameter vector $\alpha = (\beta/\sigma)$ (first brackets) and a truncated-at-zero normal distribution with parameters (β, σ) and $E(Y_i) = X_i \beta + \sigma \phi_i/\Phi_i$ (second brackets). As such, separate maximization subject to the restrictions that the probit parameter vector be a positive scalar multiple (specifically $1/\sigma$) of the parameter vector of the truncated normal model yields the Tobit model. The probit component can, of course, be viewed as the model of whether or not the threshold is crossed, while the truncated normal component models the conditional phenomenon of the magnitude of the activity given that the activity is undertaken.

It is certainly reasonable to consider the possibility that the parameter restrictions described in the preceeding paragraph are in fact invalid. This would indicate, therefore, that the model of threshold crossing is not as intimately related to the conditional model of the magnitude of the undertaken activity as is implied by the Tobit model. In the context of health outcomes, this could mean that the phenomenon of whether some illness occurs is governed by a set of parameters different than that determining the amount, duration, or severity of the illness, given that some illness occurs. We discuss such issues in greater detail later in the Chapter.

Another characteristic of the Tobit model that merits discussion is the fact that the parameters estimated under the assumptions of the Tobit model are in general nonrobust to departures from many of the underlying assumptions. That is, violation in the data of some of the properties implied when the likelihood function is written in the form (5) will lead to inconsistent estimates of the parameters (β, σ). This phenomenon, which is not uncommon in many types of models that are estimated by means of maximum likelihood, stands in contrast to more familiar formulations such as OLS and nonlinear least squares where, in spite of a variety of departures from the assumed ideal structure of the error terms, one can still obtain consistent estimates of the structural parameters.

Two of the most often discussed violations that bode dire consequences for Tobit parameter estimates are violations of the NID assumption: first, that the error variances are nonconstant across observations, and second, that the error structure, though perhaps homoscedastic, is nonnormal. Note that normal, homoscedastic errors are implied when writing the likelihood function in the form (5). The results of several studies, summarized by Amemiya (1984), suggest that under either type of departure, the maximum likelihood Tobit parameter estimates are inconsistent.

2.4 CRAGG-CLASS HEALTH OUTCOME MODELS

In a 1971 paper, Cragg proposed a set of models for situations that can be depicted as follows. An economic agent makes two (simultaneous) decisions. A dichotomous decision is made about whether or not to engage in some activity. Conditional on an affirmative for this decision, a decision is made regarding how much of the activity to pursue. The activities can be construed in the broadest of terms: expenditures, quantities demanded or supplied, or the amount of time spent in ill health. Such models have come to be known as hurdles models, that is, conditional or some hurdle being crossed, a decision is made about some magnitude of interest. Although these processes might in some cases seem logically to be ordered in a temporal manner, the statistical properties of the model abstract from any temporal considerations, the quantity decision being described in terms of conditional densities.

Cragg proposed several models. However, because of the nature of the present study, only two members of this set will concern us here, these being the formulations wherein the quantity or second-stage decision is defined only on the positive reals. This is in obvious reference to ideas like "given that an individual had some illness, how much time was spent ill." Although Cragg's other formulations are also interesting, their discussion is omitted for economy of space.

For notational ease, we will assume that the same vector of independent variables, X_i, influences both the first-and second-stage decisions. This is a completely innocuous assumption, however, as elements of parameter vectors can be restricted equal to zero to accommodate more general cases. Regardless of the specification of the second-stage or conditional decision, the first-stage is described by a binary probit model, i.e. the existence of latent random variables $Y^*_{i1} \sim N(X_i\beta_1, \sigma_1^2)$ is posited. Only the signs of the realizations are recorded, however, and are codified according to

$$y_{i1} = 1, \; y^*_{i1} \geq 0 \qquad (9)$$

$$= 0, \; y^*_{i1} < 0$$

Because of this codification scheme, there is no information about the scale of the random variables Y^*_{i1} (i.e. the mappings of y^*_{i1} into y_{i1} are unaffected by transformations of Y^*_{i1} of the form θY^*_{i1} for $\theta > 0$). Therefore, some normalization is required, the most common being $\sigma_1 = 1$. This formulation gives rise to Cragg's formulation of the hurdle-crossing model, where, with obvious change from Cragg's notation, we specify

$$\Pr(y_{i1} = 1) = \Phi(X_i\beta_1) \qquad (10)$$

$$\Pr(y_{i1} = 0) = \Phi(-X_i\beta_1),$$

where Φ is the standard normal distribution function (Cragg uses $C(\cdot)$ for $\Phi(\cdot)$).

For strictly positive second-stage quantity realizations, Cragg proposes two alternative formulations. Both are based on the specification of the conditional densities for random variables Y_{i2} given that the activity is in fact undertaken.

The first formulation is one where the conditional density for the realizations of the Y_{i2} is truncated-normal, with the truncation point at zero. Thus we have

$$g(y_{i2}|y_{i1}=1) = \phi\left(\frac{y_{i2}-X_i\beta_2}{\sigma}\right)/\sigma\Phi\left(\frac{X_i\beta_2}{\sigma}\right), \qquad y_{i2} > 0 \qquad (11)$$

$$= 0 \qquad , \quad \text{else},$$

where ϕ and Φ are as defined earlier. With obvious notational change from Cragg's article, the (unconditional) likelihood of the positive realizations, can be written as

$$f(y_{i2}) = g(y_{i2}|y_{i1}=1)\Pr(y_{i1}=1) = \tag{12}$$

$$\phi\left(\frac{y_{i2} - X_i\beta_2}{\sigma}\right)\Phi(X_i\beta_1)/\sigma\Phi\left(\frac{X_i\beta_2}{\sigma}\right)$$

for $y_{i2} > 0$. Therefore, the likelihood function of Cragg's first model is

$$L = \prod_{i\varepsilon\Omega_0} \Phi(-X_i\beta) \prod_{i\varepsilon\Omega_1} \phi\left(\frac{y_{i2} - X_i\beta}{\sigma}\right)^2 \Phi(X_i\beta_1)/\sigma\Phi\left(\frac{X_i\beta_2}{\sigma}\right) \tag{13}$$

where Ω_0 is the index set for i such that $y_{i1} = 0$ and Ω_1 is the index set for $y_{i1} = 1$. Written in log form,

$$\ell = \sum_{i\varepsilon\Omega_0} \ln(1-\Phi(X_i\beta_1)) + \sum_{i\varepsilon\Omega_1} \ln\phi\left(\frac{y_{i2} - X_i\beta_2}{\sigma}\right) + \ln\Phi(X_i\beta_1) - \ln\sigma - \ln\Phi\left(\frac{X_i\beta_2}{\sigma}\right). \tag{14}$$

In the form (14), it is straightforward to see that maximization of ℓ is fully equivalent to the two-stage maximization problem:

1) Probit estimation of the parameter vector β_1 via maximization of

$$\ell_1 = \sum_{i\varepsilon\Omega_0}(\ln(1-\Phi(X_i\beta_1)) + \sum_{i\varepsilon\Omega_1}\ln\Phi(X_i\beta_1); \tag{15}$$

2) Truncated-normal estimation of the parameters (β_2, σ) via maximization of

$$\ell_2 = \sum_{i\varepsilon\Omega_1} \ln\phi\left(\frac{y_{i2} - X_i\beta_2}{\sigma}\right) - \ln\sigma - \ln\Phi\left(\frac{X_i\beta_2}{\sigma}\right). \tag{16}$$

Because of the complexity of the log likelihood (14), estimation in this two-stage fashion is likely to be somewhat easier than attempting to maximize (14) with respect to the (2k+1) parameters $(\beta_1, \beta_2, \sigma)$.

Cragg's second formulation again depends on the probit first-stage model, but the conditional density of the positive realizations is respecified. Instead of assuming that the conditional density of the positive realizations of Y_{12} is truncated-normal, the model is now formulated such that the logarithms of the y_{i2} are normal, i.e. conditional on $y_{i1}=0$, $\log(y_{i2}) \sim N(X_i\beta_2, \sigma^2)$. The conditional density for the $i\epsilon\Omega_1$ is

$$h(y_{i2}|y_{i1}=1) = (y_{i2}\sigma)^{-1}\phi\left(\frac{\log(y_{i2})-X_i\beta_2}{\sigma}\right) \qquad (17)$$

where the term $(y_{i2})^{-1}$ is the Jacobian of the transformation from y_{i2} to $\log(y_{i2})$. Therefore, the likelihood for the $i\epsilon\Omega_1$, which is Cragg's equation (11), is

$$f(y_{i2}) = h(y_{i2}|y_{i1}=1)\Pr(y_{i1}=1)$$

$$= (y_{i2}\sigma)^{-1}\phi\left(\frac{\log(y_{i2})-X_i\beta_2}{\sigma}\right)\Phi(X_i\beta_1) \qquad (18)$$

The likelihood function for the entire sample is

$$L = \prod_{i\epsilon\Omega_0}\Phi(-X_i\beta_1)\prod_{i\epsilon\Omega_1}(y_{i2}\sigma)^{-1}\phi\left(\frac{\log(y_{i2})-X_i\beta_2}{\sigma}\right)\Phi(X_i\beta_1) \qquad (19)$$

In log form,

$$\ell = \sum_{i\epsilon\Omega_0}\ln(1-\Phi(X_i\beta_1)) + \sum_{i\epsilon\Omega_1}\ln\phi\left(\frac{\log(y_{i2})-X_i\beta_2}{\sigma}\right) + \ln\Phi(X_i\beta_1)$$

$$-\ln y_{i2} - \ln\sigma \qquad (20)$$

As in Cragg's first model, the second model can be estimated in two stages:

1) Probit estimation of β_1 as above;

2) OLS estimation of (β_2, σ) using the log transform of the y_{i2} as dependent variables and X_i as the independent variables. This is perhaps surprising, but results because the terms in (20) involving (β_2, σ) are identical to those of the likelihood function of the familiar normal linear model.

Because of the simplicity of this two-stage approach, estimation in such a framework is obviously appealing. Duan, et. al. (1983) have proposed the second Cragg model to estimate medical expenditures: individuals either have or do not have medical expenses, and given that they have medical expenses, the conditional density of the expenditures is lognormal, $\log(Y_{i2}) \sim (X_i\beta_2, \sigma^2)$. However, the Duan, et. al. model likelihood function is incorrect. They omit the multiplicative term $(1/y_{i2})$ in the density for the positive terms, this as mentioned above being the Jacobian of the transformation from y_{i2} to $\log(y_{i2})$. Cragg correctly incorporates this in his exposition. The values of the parameters that maximize the log-likelihood function do not depend on this transform, although the value of the log-likelihood function itself does, of course, depend on the transform.

2.5 TRUNCATED-NORMAL ESTIMATION

As described above, estimation of the truncated-normal model is the relevant second step in estimating the first Cragg model where the positive observations are assumed to follow a truncated-from-below normal distribution. Although there are several variants of the truncated normal (including truncated-from-below, truncated-from-above, doubly-truncated, and constant or nonconstant point(s) of truncation), the discussion here will concentrate on the case most relevant to the present empirical work, viz. the truncated-from-below distribution where the point of truncation is constant across observations and is assumed to be zero. The results easily generalize, however, and for a discussion of the statistical properties of the truncated normal distribution in the most general case, the reader is referred to Johnson and Kotz (1970, pp. 81-87).

It should be noted that interest in the truncated normal should not be confined to the role it plays in the Cragg model. The distribution is indeed useful in many empirical situations. Hurd (1979) notes that

> (e)stimation based on only positive y's comes about very naturally in a number of kinds of studies. For example, in many labor supply studies one of the right-hand variables, the wage rate, is only observed when the left-hand variable, labor supply, is positive. Imputing the unobserved wage rates causes a number of complications that can be avoided by discarding those observations for which labor supply is zero. Another example is a demand study where the price is not known unless a purchase is made. (Hurd, 1979, p. 248).

For our purposes, the likelihood function of the truncated normal can be constructed as follows. We assume the existence of $T_1 + T_2$ realizations of random variables Y_i-NID $(X_i\beta, \sigma^2)$. However, for whatever reasons, only the positive realization of the Y_i are used in the analysis, these assumed to number T_1. Given these assumptions, the likelihood function is

$$L = \prod_{i=1}^{T_1} (\phi_i / \sigma\Phi_i), \tag{21}$$

where ϕ_i is the standard normal density evaluated at $((y_i - X_i\beta)/\sigma)$, and Φ_i is the standard normal distribution function evaluated at $(X_i\beta/\sigma)$ which serves as the normalizing factor of the truncated density. The log-likelihood function (suppressing terms not depending on (β,σ)) is

$$\ell = \sum_{i=1}^{T_1} -.5((y_i - X_i\beta)/\sigma)^2 - \log\sigma - \log\Phi(\frac{X_i\beta}{\sigma}) \tag{22}$$

Estimation is by means of maximum likelihood. The first-order conditions for a maximum of ℓ are

$$\begin{aligned} \partial\ell/\partial\beta &= \sum_{i=1}^{T_1} [\frac{-\phi_i}{\sigma\Phi_i} + (\frac{y_i - X_i\beta}{\sigma^2})]\, X_i' = 0 \\ \partial\ell/\partial\sigma &= \sum_{i=1}^{T_1} (\frac{\phi_i}{\Phi_i})(\frac{X_i\beta}{\sigma^2}) - \frac{1}{\sigma} + \frac{(y_i - X_i\beta)^2}{\sigma^3} . \end{aligned} \tag{23}$$

The second derivatives are complicated and will not be presented here. Experience has demonstrated that the Berndt-Hall-Hall-Hausman first derivative approach for optimization works rather well.

In the case of the truncated-normal, like most all other limited dependent variable models, ordinary least squares estimation of the parameters (β,σ) yields biased and inconsistent estimates. However, in the case of the truncated normal, Olsen (1980) has shown how the OLS estimates can be used fruitfully to generate estimates of (β,σ) that, while inconsistent, can provide remarkably good approximations to the maximum likelihood estimates (our experience in other areas is consistent with Olsen's finding) and, as such, serve as excellent starting values for maximum likelihood estimation algorithms.

Olsen's method relies on a method of moments technique whereby the moments (specifically the mean and variance) of the empirical incomplete distribution that of the positive y_i, are related to the moments of the complete distribution via formulae developed by Pearson and Lee (1908). Extending the Pearson-Lee methodology to the multiple regression case, Olsen demonstrates that the least squares slope coefficients differ from the true slope coefficients by a common factor, and he presents in tabular form the multiplicative correction factors needed to transform the OLS estimates of the slope, intercept, and standard error parameters (based on data from the incomplete distribution) to the corresponding complete distribution estimates. In practice, we have fitted polynominal functions of the third degree to Olsen's tabled data so that the transformations are facilitated.

Olsen also presents the multipliers for transforming the (mean/standard error) ratio estimated by OLS on the incomplete distribution to the corresponding ratio of the complete distribution, (μ/σ). Olsen notes that $\Phi(\mu/\sigma)$, where Φ is the standard normal cumulative distribution, should give an idea of the expected ratio of nontruncated to total observations. Therefore, if one is considering Tobit estimation of the parameters of a censored distribution, it should hold that the $\Phi(\mu/\sigma)$ based on the (μ/σ) estimated using Olsen's method and treating the nonlimit observations as truncated normal should accord approximately with the ratio of noncensored to total observations. There is no formal test to assess how closely these should accord however. Olsen suggests that a disagreement here could well indicate that the Tobit is an inappropriate specification.

2.6 HECKMAN'S APPROACH: SAMPLE SELECTION

A very popular technique for estimating models with limited dependent variable estimation is the sample selection model, attributable largely to Heckman (1976, 1979). The model has a number of applications (see Heckman's 1976 article in particular), and is quite easy to estimate. Because it is so well-known, we will only provide a sketch of the details. The following section, which contrasts and compares the Tobit, Cragg, and Heckman models, sheds some more light on subtleties of Heckman's formulation.

Heckman considers the following two-equation model:

$$y_{i1} = X_1\beta_1 + \varepsilon_{i1} \tag{24}$$

$$y^*_{i2} = X_1\beta_2 + \varepsilon_{i2} \tag{25}$$

It is assumed that ε_{i1} and ε_{i2} are distributed joint normal, with marginal densities $N(0, \sigma_1^2)$ and $N(0, \sigma_2^2)$ respectively, and covariance σ_{12}. It can be further assumed that the realizations y^*_{i2} are unobserved. However, available are discrete sign indicators y_{i2} which are mapped as

$$\begin{aligned} y_{i2} &= 1, \; y^*_{i2} > 0 \\ &= 0, \; y^*_{i2} \leq 0. \end{aligned} \tag{26}$$

In Heckman's model, the realizations y_{i1} are available to the analyst only when $y^*_{i2} > 0$, i.e. when $y_{i2} = 1$.

A concrete example is where (24) is a model determining market wage rate (or log(wage rate)) by a linear function of X_i and random error and where (25) is a model determining hours of labor supplied in the market. It is assumed that either hours of labor supplied or a discrete binary indicator of whether or not any hours were supplied is available for all observations. However, because market wage rates are only observed for individuals for whom the market wage rate exceeds the reservation wage at zero hours, data on the y_{i1} are available only when $y_{i2}^{*} > 0$ $(y_{i2} = 1)$.

Heckman then considers the expectation $E(Y_{i1}|y_{i2} = 1)$, which can be written as

$$E(Y_{i1}|y_{i2} = 1) = X_i\beta_1 + E(\varepsilon_{i1}|y_{i2} = 1). \tag{27}$$

If one considers least-squares estimation of (27), the question is: Do there obtain consistent estimates of β_1 when y_{i1} is regressed on X_i for those i for whom $y_{i2}=1$? Basically the issue is whether the expectation $E(\varepsilon_{i1}|y_{i2} = 1)$ is null. In general, and thus at the core of the sample selection bias problem, the answer is "no". Based on well-known formulae, it holds that

$$E(\varepsilon_{i1}|y_{i2} = 1) = \sigma_{12}\phi_i/\sigma_2(1-\Phi_i), \tag{28}$$

where ϕ_i is the standard normal density evaluated at $(X_i\beta_2/\sigma_2)$ and Φ_i is the distribution function evaluated at the same point. Because σ_{12} is in general nonzero and since ϕ_i, $(1-\Phi_i)$, and σ_2 are all positive, then least squares estimation of (27) will be based on an expectations function with nonnull disturbance expectation, and will therefore yield inconsistent estimates of β_1.

Heckman's suggested procedure in this situation is as follows. Estimate on the entire sample a probit model for the discrete indicator representation of the model (25). This yields a consistent estimate of the parameter vector (β_2/σ_2) from which consistent estimates of $\lambda_i = \phi_i/(1-\Phi_i)$ are constructed. Form the Tx(k+1) matrix $Z = [X|\Lambda]$, where Λ is a Tx1 vector with typical element λ_i, and regress y_i on $[x_i, \lambda_i]$. This procedure yields consistent estimates of the parameters β_1 and (σ_{12}/σ_2), having effectively solved the omitted variables problem by using a consistent estimate of $E(\varepsilon_{i1}|y_{i2} = 1)$ as a regressor.

In the context of health outcomes models, one could define y^*_{i2} as some latent index of the propensity to be ill. Given that this index is greater than some threshold level, illness results, its magnitude determined by the realization y_{i1}. The translation of the latent illness model into Heckman's framework is not straightforward, however. For those individuals not reporting illness over the sample interval, we observe zero time spent ill rather than not observing the amount. It is therefore difficult to interpret the meaning of the realized, but unobserved, y_{i1} for the healthy. We turn in the next section to a more detailed analysis of such subtleties.

2.7 TOBIN, CRAGG, AND HECKMAN: A DIGRESSION

As there are some similarities between and among the models described above and identified for expositional parsimony as the models of Tobin, Cragg, and Heckman, it is appropriate to summarize their similarities and differences and in so doing to elucidate the circumstances in which each model is more or less appropriate. (The discussion of Cragg's model here is the second Cragg model (probit/truncated-normal), as that version is most similar to the others discussed here.)

First to note is that the Tobit model results as a restricted version of both the Cragg and the Heckman models. The reason for this is purely mechanical, however, and should not be taken to imply that the Cragg and Heckman models are in general identical. As we will see below, these models are structurally quite different.

To see that the Cragg model reduces to the Tobit, the Cragg log-likelihood function can be written (following Lin and Schmidt (LS) (1984)) as

$$\ell = \sum_{i\in\Omega_0} \ln\Phi(-X_i\beta_1) + \sum_{i\in\Omega_1} \Big[\ln\Phi(X_i\beta_1) - \ln\Phi\Big(\frac{X_i\beta_2}{\sigma}\Big) - (1/2)\ln(2\Pi\sigma^2) - (1/2\sigma^2)(y_i - X_i\beta_2)^2\Big] \tag{29}$$

If the restriction $\beta_1 = \beta_2/\sigma$ is imposed, then the first two terms in the square brackets cancel and (29) is easily seen to be identical to (6) with β in (6) replaced by β_2 from (29). The upshot of such parameter restrictions is hardly trivial, however. As specified, and discussed briefly earlier, the Tobit model is fairly restrictive in its behavioral implications, as the parameter vector that governs the probability of observing an above-threshold realization of the dependent variable is the same as that governing the quantity realization of the dependent variable given that it is above the threshold. Owing to the implications of such restrictions, LS have concluded that "...the Tobit model is typically used with more faith than it warrants..." and have developed a test (which we discuss below) for the appropriateness of the $\beta_1 = \beta_2/\sigma$ restriction of the Cragg model that is implied by the Tobit specification. The following excerpt from LS provides a particularly cogent summary description of the appropriateness of the restricted (Tobit) versus the unrestricted versions of the Cragg model:

> (I)n the Tobit model any variable which increases the probability of a non-zero value must also increase the mean of the positive values; a positive element of β means that an increase in the corresponding variable (element of X_i) increases both $\Pr(y_1>0)$ and $E(y_1|y_1 > 0)$. This is not always reasonable. As an example, consider a hypothetical sample of buildings, and suppose that we wish to analyze the dependent variable "loss due to fire," during some time period. Since this is often zero but otherwise positive, the Tobit model might be an obvious choice. However, it is not hard to imagine that newer (and more valuable) buildings might be less likely to have fires, but might have greater average losses when a fire did occur. The Tobit model can not accommodate this possibility.
>
> Another problem with the Tobit model is that it links the shape of the distribution of the positive observations and the probability of a positive observation. For rare events (like fires), the shape of the distribution of the positive observations would have to resemble the extreme upper tail of a normal, which would imply a continuous and faster than exponential decline in density as one moved away from zero. Conversely, when zero occurs less than half of the time, the Tobit model necessarily implies a non-zero mode for the non-zero observations.
>
> Cragg's model avoids both of the above problems with the Tobit model. A reasonably strong case can be made for it as a general alternative to the Tobit model, for analysis of data sets to which Tobit is typically applied—namely, data sets in which zero is a common (and meaningful) value of the dependent variable, and the non-zero observations are all positive. The distribution of such a dependent variable is characterized by the probability that it equals zero and by the (conditional) distribution of the positive observations, both of which Cragg's model parameterizes in a general way. (LS, pp. 174-175)

Turning now to Heckman's formulation, his two-equation model is seen to reduce to the Tobit model as follows. Recall that the model can be written (with notational changes obvious) as

$$y_{i1} = X_i\beta_1 + \varepsilon_{i1} \tag{30}$$

$$y^*_{i2} = X_i\beta_2 + \epsilon_{i2}.$$

Y^*_{i2} is a latent variable, however, and only a discrete (0,1) sign indicator of its realization y_{i2} is available. y_{i1} is observed only when $y_{i2} = 1$. Letting $\beta_1 = \beta_2$ and $\epsilon_{i1} = \epsilon_{i2}$ (i.e. the error structure is univariate rather than bivariate), then the Heckman model is the standard Tobit model. The logic is that when these restrictions are imposed in the Heckman two-equation model, the remaining single equation plays both the censoring and the determination-of-intensity roles. Since the censoring occurs as a result of a non-positive realization of the random variable Y^*_{i2}, the Tobit requirement that the quantity or intensity realization be confined to the nonnegative orthant is automatically satisfied when the restriction $y_{i1} = y^*_{i2}$ (i.e. $\beta_1 = \beta_2$, $\epsilon_{i1} = \epsilon_{i2}$) is imposed. In general, however, the Heckman two-equation framework is not specifically designed to model situations where realizations of the dependent variable of interest are necessarily nonnegative and are recorded for all individuals/observations, and where $Pr(y_i = 0) > 0$. Heckman's formulation has $y_i \neq 0$ except on a set of measure zero. We turn now to an explanation of the fundamental differences between the Heckman two-equation formulation and the two versions-of-interest of the Cragg model.

The two-equation Heckman model describes two phenomenon, Y_{i1} and Y^*_{i2}, that are marginally distributed, respectively, as $NID(X_i\beta_1, \sigma^2)$ and $NID(X_i\beta_2, \sigma^2_2)$ (σ^2_2 is usually restricted = 1 for normalization when only the sign of y^*_{i2} is observed). The joint distribution is bivariate $NID(X_i\beta_1, X_i\beta_2, \sigma^2_1, \sigma^2_1, \rho)$, where ρ is the correlation of $(\epsilon_{i1}, \epsilon_{i2})$, $(\sigma_{12}/\sigma_1\sigma_2)$, which is in general nonzero. The important point is that these marginal and joint distributions are unconditional. That is, for all i, there exist realizations (y_{i1}, y_{i2}^*) although the realizations y_{i1} for some i will be unavailable to the researcher. Casting the problem concretely in the area where Heckman's model has been most fruitfully applied, labor economics, sheds further light on the subtleties of his model. Here we define $y_{i1}=\log(W_i)$ and $y_{i2}^*=\log(H_i+1)$, where W_t is wage earned in market work and H_i is hours of market work. Thus, y_{i2}^* is positive only if market hours are positive. It is posited that the expected values of both Y_{i1} and Y_{i2} are linear functions of personal characteristics and other variables so that the two-equation model results. However, because we only observe the market wage for those individuals actually participating in market work (those for whom $H_i>0$), some subset of observations will not have data on the y_{i1}. There is a market wage determined for nonparticipants; whether or not such individuals have knowledge of their market wages is immaterial. The relevant analytical fact is that such data are unavailable to the researcher.

In this labor supply framework, it is apparent why the estimation techniques developed for the two-equation Heckman model and discussed earlier in this chapter have such appeal. The more immediate concern, of course, is whether such techniques are in fact appropriate to the estimation requirements of the present analysis. In a nutshell, Heckman's model is one where there are two equations of interest, both holding for all i unconditionally, and where (except when restricted so as to be identical to a Tobit model) the probability of observing realizations of the dependent variable equal zero is zero. Does such a formulation capture the essence of the "corner solution" problems of the health status outcomes phenomenon?

It seems rather artificial to cast such phenomena in such this framework. It is not generally the case with the generation of health outcomes data that one can posit the existence of some latent variable such that data for the illness measure(s) of interest are only available given a positive realization of the latent variable. Rather, the processes of interest here are represented more typically by data that indicate the realizations of illness outcomes for all individuals, even though these realizations are quite frequently on the boundary of the "consumption" set. In sketching some of the differences between the Heckman two-equation formulation and the Cragg models with particular reference to data sets where the zero or corner solution outcomes are meaningful and where nonzero outcomes are strictly positive, LS observe that in such cases the Heckman model's assumptions are not particularly representative of the situation because in the Heckman formulation:

> ... the observed values of y_{i1} need not be positive, in the sense that the model implies a non-zero probability of observed $y_{i1} < 0$; and the unobserved y_{i1} are literally unobserved, rather than observed as equal to zero. The first of these problems can be circumvented, for example, by measuring y_{i1} in logarithms,...and the second problem is in any case fundamental. (LS p. 175).

We turn now to a discussion of how the Cragg models differ in substance from the Heckman two-equation setup and argue that the Cragg formulations are relatively more suited than Heckman's model to the nature of a subset of our estimation requirements.

Although like the Heckman formulation in being a "two-model" specification, the fundamental point of departure for the Cragg technique is that one of the two models is formulated in terms of conditional expectations. The conditions on which the expectations are taken are, as described above, the outcomes of unconditional models, which are generally stated as binary representations of latent random variables. Thus, in the context of health measures, there is an unconditional model defined for all individuals determining the binary outcome (illness, no illness). Conditional on an "illness" outcome, the quantity or duration of illness is determined either by a lognormal or truncated-normal model. The unconditional likelihood for a representative ill individual is then

density(illness duration given some illness)*Pr(some illness), (31)

which is equation (12) as specified earlier. There is no density of the quantity of illness defined for the healthy, unlike Heckman's formulation that defines such a density for all individuals.

Deaton and Irish (DI) (1984), in an independent line of investigation, have purportedly cast Cragg's first model in a two-equation Heckman formulation. They indicate that a positive observation on the quantity measure of interest is made when, in the notation used earlier, both Y_{i1} and Y^{*}_{i2} are realized as positive, else a zero or a nonparticipation results. In two cases, DI specify

$$\begin{aligned} y_{i1} &= X_i \beta_1 + \varepsilon_{i1} \\ y^{*}_{i} &= X_i \beta_2 + \varepsilon_{i2} \end{aligned} \qquad (32)$$

Cast thusly, the Cragg model can be viewed as a Heckman two-equation model, but with a restriction imposed that is absent in Heckman's formulations. That is DI seem to have ignored one aspect of the Cragg model that is key in differentiating it from Heckman's specification, viz. that $y_{12}^{*}>0$ is both a necessary and sufficient condition for a positive realization of y_{i1} to result. That is, $\Pr(y_{i1}>0|y_{i2}^{*}>0) = 1$, $\Pr(y_{i1}=0|y_{i2}^{*}<0) = 1$. When, and only when the first hurdle is traversed is there a positive amount of the activity undertaken. So DI's statement that positive realizations of both variables determines whether y_{i1} is observed positive is somewhat misleading in that a positive realization of either suffices to assume the positivity of the other. Neither of Cragg's specifications, then, is really in the spirit of the model proposed by Heckman except, of course, when both the Cragg model and the Heckman two-equation formulation are restricted such that the Tobit specification results.

Owing to the subtleties of the arguments, it is likely that the above discussion has provided somewhat less than a total clarification of all the relevant issues. Some of these shortcomings are due to the fact that even central participants in the academic debates appear still unconvinced about the nature of the differences among the estimation techniques. For example, as noted earlier Duan and coauthors (1983) have used the Cragg estimation technique to model individuals' medical expenditures. The expenditure decision, in the spirit of Cragg's specification, is statistically modeled as two separate processes. Model one determines the binary outcome of whether or not any expenditures will occur, and model two determines the amount of expenditure (positive by definition) that results conditional on there being some expenditure. In this paper, Duan and coauthors assert that the covariance between the error terms of the two models is irrelevant insofar as construction of the likelihood function is concerned.

Recently, however, Hay and Olsen (1984) have questioned the Duan and coauthors method, stating that this approach "requires some fairly unusual assumptions on the model joint error distribution and functional form (p. 279)." Moreover, Hay and Olsen go on to claim that the Duan and coauthors formulation "can be interpreted as being nested in the more general sample selection models (p. 279)." Duan and coauthors respond that Hay and Olsen "are incorrect in claiming that our models are nested within the sample selection model," and that "the conditional specification in the multi-part (i.e., Duan and coauthors) model is preferable to the unconditional specification in the selection model for modeling actual (v. potential) outcomes (p. 283)."

As we argued earlier, the sample selection or Heckman approach is particularly fruitful when analyzing phenomena such as labor market participation. Quoting Duan and coauthors:

> For certain empirical problems such as labor force participation, the primary goal might be to predict the potential outcome instead of the actual outcome; therefore, an unconditional specification such as the sample selection models might be preferable. For the present application, however, the goal is to predict the actual expense, not the potential expense; therefore, the unconditional equation... is of no direct interest, and the preference for the unconditional specification in the other empirical problems does not apply to the present application. (p. 286).

In any event, this discussion demonstrates that there still exists some confusion on these points in the published literature. We have attempted to be as thorough as time and space permit in hope of emphasizing one extremely important message. That is, it is essential that the researcher be intimately familiar with the behavioral and statistical structure of the models of interest in order to avoid being swallowed by the slippery quicksand we have described. The nature of health status measures as conditional or unconditional and the interpretation of any latent variables in the model must be quite clear before the correct estimation technique can be selected. When, and only when, such issues are in order is it possible to make sense of the estimated obtained and their relevance to benefit estimation.

It seems that the logic of the health status outcome measures of interest in this study is better captured in terms of Cragg's specifications than in the Heckman two-equation model although this question is obviously still open to informed debate. The specification of the magnitude-of-illness model as a conditional model is, however, intuitively plausible, and Cragg's formulations provide a natural vehicle for translating such intuitive plausibility into an econometric framework. However, it so happens that the assumption of normally, or at least continuously, distributed random variables, which characterizes the above models, is not necessarily appropriate insofar as count measures like "days" or "times" are concerned. To a discussion of some alternative estimation techniques that might be used in such situations we now turn.

2.8 POISSON-DISTRIBUTED HEALTH OUTCOME MEASURES

In modeling event counts (non-negative integer data) over some time interval (t, t+dt), the Poisson distribution is commonly used. Here, a random variable Y_i follows the probability law

$$\begin{aligned} Pr(Y_i = y) &= \exp(-\lambda_i)\lambda_i^y/y!, \quad y\epsilon\{0,1,2,\ldots\} \\ &= 0 \qquad\qquad\qquad\quad , \text{ else} \end{aligned} \tag{33}$$

43<

with $E(Y_i) = Var(Y_i) = \lambda_i$.

It happens that there exist health outcome data of interest that are recorded as nonnegative integers, most obviously as counts of days of activity restriction. For any individual, such measures can, over a time interval (t, t+dt), say one two-week period, assume only integer values in $\{0,1,2,\ldots,14\}$. Because of the paucity of observations likely to be found at the upper (14 day) limit, we ignore the fact that these measures obey upper bounds and concentrate instead on the complications presented by the large number of individuals who in a typical random sample of the population report zero days of restricted activity.

Analogous to the familiar normal distribution where for econometric work one typically specifies $\mu_i = X_i\beta$, the λ_i parameter of the Poisson distribution can be reparameterized to admit the influence of covariates. Since for all i, $\lambda_i > 0$, a straightforward approach is to assume $\lambda_i = \exp(X_i\beta)$ and to estimate β by maximum likelihood (see Hausman, Hall, Griliches (1984), Hausman, Ostro, Wise (1983), Portney and Mullahy (1985)). This is the approach adopted here for modeling the restricted activity day outcomes.

One drawback of the Poisson model is the restriction that $E(Y_i) = Var(Y_i)$. Should this restriction not in fact characterize the data, the maximum likelihood estimates of the covariance matrix of $\hat{\beta}$ based on minus the inverse of the estimated Hessian will be inconsistent and t-tests based thereon would be misleading. Hausman, Ostro, and Wise circumvent this restriction by allowing for an overdispersion parameter. A different approach is used here, using an estimator of the covariance matrix that is robust against departures from the mean-variance restriction, this procedure described below.

Given T independent observations, the log-likelihood function of the Poisson health outcome model can be written as

$$\ell = \sum_i -\exp(X_i\beta) + y_iX_i\beta + C, \qquad (34)$$

where $\exp(X_i\beta) = \lambda_i$, y_i is the observed count of illness days, and C does not depend on β. It is obvious that ℓ is concave in β. The first-order conditions for the maximization of ℓ are

$$\partial\ell/\partial\beta = \sum_i -\exp(X_i\beta)X_i' + y_iX_i' = 0 \qquad (35)$$

with the maximum guaranteed by the condition

$$\partial^2\ell/\partial\beta\partial\beta' = \sum_i -(X_i'X_i)\exp(X_i\beta) \qquad (36)$$

negative definite.

The maximum likelihood estimates of β obtained by maximizing (46) are consistent, but the estimate of the covariance matrix of $\hat{\beta}_{ML}$ using $[-\partial^2\ell/\partial\beta\partial\beta']^{-1}$ evaluated at $\hat{\beta}_{ML}$ will be inconsistent if the data are not in fact generated by the specified Poisson distribution.

This is most easily seen as follows. Note that the model can be equivalently cast as a nonlinear least squares regression, the i-th observation being

$$\begin{aligned} y_i &= E(Y_i) + \varepsilon_i \\ &= \exp(X_i\beta) + \varepsilon_i \end{aligned} \qquad (37)$$

with $E(\varepsilon_i) = 0$. Clearly, $\text{var}(\varepsilon_i) = \text{var}(Y_i) = \exp(X_i\beta)$, so that the ε_i are heteroscedastic. If nonlinear weighted least squares is used with the weights $\exp(-X_i\beta)$ formed using consistent estimates of β, and if the data are in fact Poisson as specified, the maximum likelihood consistent estimates of β and $\mathrm{cov}(\hat{\beta})$ will obtain. (The consistency of $\hat{\beta}_{ML}$ for β does not depend on the weighting scheme.) However, if the data is not Poisson-distributed, the estimate of $\mathrm{cov}(\hat{\beta})$ obtained in this manner will be inconsistent and asymptotic t-tests based thereon will be misleading. The case is fully analogous to the estimation of the heteroscedastic linear model which yields inconsistent covariance estimates (and, therefore, t-statistics) if the heteroscedastic nature of the error structure is either ignored or incorrectly specified.

Royall (1984) has demonstrated a method whereby estimates of $\mathrm{cov}(\hat{\beta})$ robust against misspecification of the underlying distribution of the data can be obtained for various distributions, including the Poisson, when $[-\partial^2 \ell/\partial\beta\partial\beta']^{-1}$ evaluated at $\hat{\beta}_{ML}$ fails to yield a consistent estimate of $\mathrm{cov}(\hat{\beta})$. Denoting $I(\beta)$ as $[-\partial^2 \ell/\partial\beta\partial\beta']$, Royall's suggestion is to estimate $\mathrm{cov}(\hat{\beta})$ as

$$I(\beta)^{-1}\Big[\sum_i(\partial\ell_i/\partial\beta)(\partial\ell_i/\partial\beta)'\Big]I(\beta)^{-1} \qquad (38)$$

where ℓ_i is the i-th observation's contribution to the log-likelihood function and where all relevant evaluations in (38) are at $\hat{\beta}_{ML}$. This will be the approach adopted in empirical implementation of the Poisson model the present study.

2.9 GEOMETRIC-DISTRIBUTED HEALTH OUTCOME MEASURES

One alternative to the Poisson model for the modeling of count data is the geometric distribution. Though seemingly not as often used by econometricians as the Poisson, the geometric is a logical choice should an alternative to the Poisson be desired. Furthermore, the basic geometric specification does not suffer from the mean-variance restriction that is implied in the basic Poisson model. As will be seen below, the variance of a geometric-distrbuted discrete random variable is greater than its mean, although the fact that the variance depends on the mean limits somewhat the flexibility of the distribution.

Our description of the properties of the geometric distribution follows that of Johnson and Kotz (1969). First, it should be noted that the geometric is a special case of the negative binomial. Discussion is confined here to the geometric because it is computationally far more straightforward than is the general negative binomial. The geometric distribution is defined as follows:

$$\begin{aligned} \Pr(X=k) &= P^{k}(1+P)^{-(k+1)}, \quad k = 0,1,2\ldots \\ &= 0 \quad\quad\quad\quad\quad\quad\quad\quad, \text{ else} \end{aligned} \tag{39}$$

with $P>0$. It holds that $E(X) = P$ and $Var(X) = P(1+P)$. As in the econometric specification of the Poisson model considered earlier, one allows the P to vary across observations as P_i, and again $P_i = \exp(X_i\beta)$ is a sensible parameterization due to the required positivity of the P_i.

Given this, the likelihood function for T independent observations can be written as

$$L = \prod_{i=1}^{T} \exp\,(k_i X_i \beta)(1 + \exp(X_i\beta))^{-(k_i+1)} \tag{40}$$

with loglihood

$$\ell = \sum^{T} k_i x_i \beta - (k_i+1)\log\,(1 + \exp(x_i\beta)) \tag{41}$$

$i=1$

where k_i is the observed count for the i-th observation. The ML estimate $\hat{\beta}$ satisfies

$$\partial \ell/\partial \beta = \sum_{i=1}^{T}[k_i - (k_i + 1)\exp(X_i\beta)/(1 + \exp(X_i\beta))]X_i' = 0 \qquad (42)$$

The Hessian is

$$H = \partial^2 \ell/\partial\beta\partial\beta' = \sum_{i=1}^{T} -(k_i + 1)[\exp(X_i\beta)/(1 + \exp(X_i\beta))^2]X_i'X_i, \qquad (43)$$

which is seen by inspection to be negative definite. Because it is a fairly uncluttered expression, estimation and inference can proceed using $-H$ as an estimate of the information matrix and $(-H)^{-1}$ as an estimate of the covariance matrix. Unfortunately, much like the Poisson specification, the covariance estimate thus obtained is not robust to departures from the data being in fact geometric. However, the methods proposed by Royall (1984) and described for the Poisson model can be used for the geometric distribution also. As the development is identical, the details are omitted for economy of space.

2.10 MULTINOMIAL-DISTRIBUTED HEALTH OUTCOME MEASURES

One type of micro data of particular interest in health econometrics is of the following nature. We observe over the course of some fixed time period (say one two-week period) the number of times (say days) that an individual's health status is characterized by (k-1) mutually exclusive illness outcome measures and, therefore, the number of days on which no illness resulted, which can be viewed as the k-th activity. To be concrete, the two-week illness profile for some individual who has in his/her illness "possibility set" two illnesses (minor restricted activity day (=M), and severe restricted activity day (=S)), and healthy days (=H) (=14-M-S) might look like

H = 11

M = 2

S = 1

Given observations on such health outcome profiles, it is appropriate to view the data characterizing individuals' health status as realizations of multinomial random variables (see Morey (1981) for a related discussion). Recalling from discrete statistical theory, the multinomial distribution of a random variable $\underline{Y}$ with parameters $(T; P_1, \ldots, P_k)$ can be written

$$Pr(\underline{Y} = \underline{y}) = T! \prod_{j=1}^{k} (P_j^{t_j}/t_j!), \tag{44}$$

where T is the number of trials (here days), the t_j are the number of occurrences of the j-th outcome, and P_j are the probabilities that the j-th outcome will occur on a single trial. To extend the statistical model to the health status measures, we consider each daily outcome as one trial from a multinomial distribution with individual-specific parameter vector for the m-th individual $(T_m; P_{1_m}, \ldots, P_{k_m})$. Assuming $T_m = T'_m = T$ for all m, m', we henceforth drop the subscripts on the T parameters. The profile for two weeks, then, is the 14 (by assumption independent) daily trials for each individual. The econometric objective is the estimation of the P_{j_m}, i.e. estimation of the probabilities of realizing one of the k possible outcomes on a given day.

For computational simplicity, we proceed as follows. A logistic distribution for the daily outcome probabilities is assumed. Thus, the probability that the outcome is Z on any trial is

$$P_{Z_m} = \exp(X_m\beta_Z)/\sum_{j\varepsilon\Omega} \exp(X_m\beta_j) \qquad (45)$$

for $Z\varepsilon\Omega=\{M,S,H\}$. The logistic distribution assures that for all m the multinomial requirement $(\sum_{j\varepsilon\Omega} P_{j_m}=1)$ is met.

Since the probabilities (45) are unique only up to a difference in parameter vectors $(\beta_j-\beta_{j'})$, some normalization is required. The normalization most convenient and easily interpreted is $\beta_H = 0$, so that β_M and β_S are interpreted as differences between the respective illness parameter vectors and the no-illness parameter vector.

The objective, then, is estimation of the parameter vectors β_M and β_S. This is, of course, fully analogous to the widely-used multinomial logit model where a single outcome from a set of mutually exclusive outcomes is considered. In fact, that case is merely a special case of the present exposition for which $T_m = 1$ for all m.

Estimation is by means of maximum likelihood. Assuming the existence of N independent profile draws from the population, the likelihood of the data as a function of the parameters is

$$L(\beta) = \prod_{m=1}^{N} \Pr(\underline{y}_m = \underline{y}) = \prod_{m=1}^{N} T! \prod_{j\varepsilon\Omega} (P_{j_m}^{t_{j_m}}/t_{j_m}!) \qquad (46)$$

where the P_{j_m} are as defined in (57) and where Ω is the illness-type index set. In log form,

$$\ell(\beta) = \sum_{m=1}^{N} \sum_{j\in\Omega} t_{j_m} \log P_{j_m} + C, \qquad (47)$$

where C is a constant not depending on β. Given the assumed logistic probabilities, we have

$$\ell(\beta) = \sum_{m=1}^{N} \sum_{j\in\Omega} t_{j_m} [X_m\beta_j - \log (\sum_{k\in\Omega} \exp (X_m\beta_k))] + C. \qquad (48)$$

Maximizing (48) can be accomplished with only a slight modification of most existing (single-trial) multinomial logit programs.

2.11 ESTIMATION OF GROUPED DATA MODELS UNDER THE NORMALITY ASSUMPTION

There are often institutional or other constraints in the sampling or data-recording processes that have the effect of generating inexact data for research purposes. A common case is the situation where continuous measures of interest, such as the amount of time spent in ill health, are cast in the recorded micro data as grouped or interval data. We discussed above strategies that might be considered when the outcomes are recorded as "number of days" or "number of times," i.e. where the data can be viewed as realizations of discrete statistical processes rather than as discrete/integer codings of fundamentally continuous processes. In this section we concern ourselves with the situations where the underlying processes are best viewed as continuous phenomena but where the vagaries of either the sampling or data-coding procedures are such that only a finite number of intervals in which the continuous measure is defined are determined and the only data available to the analyst are indicators of the interval bounds in which the (unknown) continuous measure is realized. For example, the latent continuous measure might be "time spent ill over some time interval y (say t)," but owing to whatever reasons, all one knows is whether $t=0$, $t \varepsilon (0, 4 \text{ days}]$, $t \varepsilon (4 \text{ days}, 8 \text{ days}]$, or $t \varepsilon (8 \text{ days}, 365 \text{ days})$ (for y=one year). The purpose of this section is to present an estimating technique designed to handle such situations.

The method is based on the work of Rosett and Nelson (RN) (1975), who developed what is known as the two-limit probit estimation technique, and of Stewart (1983), who generalized the RN method to account for multi-interval data. We will, therefore, refer to the model expounded here as the RNS method. Here is posited the existence of normally-distributed random variables $Y_i^* \sim \text{NID}(X_i\beta, \sigma^2)$. The realizations y_i^* are unobserved, however. Available is the knowledge that the realization y_i^* is an element of some proper subset of **R**. More formally, partition **R** into P (≥ 2) subsets J_k, such that $\bigcup_{k=1}^{P} J_k = R$, $J_k \cap J_j = \phi$, $\forall$ k,j. The data available to the analyst are: k (such that $y_i^* \varepsilon J_k$), $\inf(J_k)$, and $\sup(J_k)$. Note that when P=2 this reduces to the binary probit model while for P=3, the RN two-limit probit model emerges.

Following Stewart, we define the p-th interval by (A_{p-1}, A_p), and set $A_0 = -\infty$, $A_P = +\infty$. Given T independent observations, the log-likelihood function of this model can be written

$$\ell = \sum_{i=1}^{T} \log\left(\Phi_{p(i)} - \Phi_{(p-1)(i)}\right) \tag{49}$$

where $\Phi_{p(i)} = \Phi(A_p - X_i\beta/\sigma)$, Φ being the cumulative distribution of the standard normal. Estimation is by maximum likelihood. The first-order conditions for maximizing ℓ are

$$\partial\ell/\partial\beta = \sum_{i=1}^{T} (\phi_{(p-1)(i)} - \phi_{p(i)})/\sigma(\Phi_{p(i)} - \Phi_{(p-1)(i)})X_i' = 0$$

and (50)

$$\partial \ell/\partial \sigma = \sum_{i=1}^{T} (\theta_{(p-1)(i)} - \theta_{p(i)}/\sigma(\Phi_{p(i)} - \Phi_{(p-1)(i)}) = 0,$$

where $\theta_{p(i)} = (A_p - X_i\beta/\sigma)(\phi(A_p - X_i\beta/\sigma))$, and $\phi(c)$ is the standard normal density $(2\pi)^{-.5}\exp(-.5c^2)$. (Note that when $P = 2$, i.e. when the model is binary probit, a parameter normalization is required. Typically $\sigma=1$ is used. This, of course, reduces the number of first order conditions from $(m+1)$ to m, where m is the dimensionality of β.) Stewart has shown how iterative least squares can be used to obtain the ML estimate. The reader is referred to his work for the details.

2.12 SUMMARY AND CONCLUSIONS

This brief survey has attempted to present an overview of several approaches to econometric estimation of air pollution - health outcomes models in situations where the distributions health outcome data are such that methods other than linear ordinary least squares are likely to be required in order to obtain consistent parameter estimates. The data used in this study are in all instances of this "nonstandard" nature. In particular, the analysis to follow concentrates on three of the types of data described in the preceeding discussion: count data, multinomial-distributed data, and discrete indicator, or (0,1), data of the probit sort. The following chapters discuss in some detail estimation of such models, and implement some of the estimation techniques presented in the analysis of this chapter.

The scope of the present analysis precludes consideration of several interesting research issues that must be placed on the menu of future research. First, the matter of severity of chronic illnesses is left untreated. It clearly is plausible that not only the presence or absence of, for example, chronic respiratory illness is related to air pollution exposures, but also that the severity of the illness - defined by some metric of severity -- is responsive to pollution exposures as well as other covariates. A second interesting issue that merits analysis in the future is the possibility that some subset of the explanatory variables used to explain health outcomes is correlated with heterogeneous individual-specific components of the unobserved equation error terms. In the present context, it might be argued that covariates such as cigarette consumption, income, labor market status, and even air pollution exposures (on this last point, see Rosenzweig and Wolpin (1984)) are possibly correlated with unobservable errors. When such heterogeneous unobservables are present and — the crux of the problem -- are correlated with observed explanatory variables, parameter estimates obtained without explicit recognition and control for this nonzero correlation are will in general be inconsistent. Some instrumental variable technique will likely be required in order to obtain consistent parameter estimates under such circumstances.

REFERENCES

Amemiya, T. 1981. "Qualitative Response Models: A Survey," Journal of Economic Literature, vol. 19, pp. 1483-1536.

_____. 1983. "Nonlinear Regression Models," in Z. Griliches and M. D. Intriligator, eds., Handbook of Econometrics, vol. 1, (Amsterdam: North-Holland).

_____. 1984. "Tobit Models: A Survey," Journal of Econometrics, vol. 24, pp. 3-61.

Berndt, E. R., B. H. Hall, R. E. Hall, and J. A. Hausman. 1974. "Estimation and Inference in Nonlinear Structural Models," Annals of Economic and Social Measurement, vol. 3, pp. 653-665.

Breusch, T. and A. R. Pagan. 1980. "The Lagrange Multiplier Test and Its Application to Model Specification in Econometrics," Review of Economic Studies, vol. 47, pp. 239-253.

Cox, D. R. and D. V. Hinkley. 1974. Theoretical Statistics (London: Chapman and Hall).

Cragg, J. G. 1971. "Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods," Econometrica, vol. 39, pp. 829-844.

Duan, N., W. G. Manning, C. N. Morris and J. P. Newhouse. 1983. "A Comparison of Alternative Models for the Demand for Medical Care," Journal of Business and Economic Statistics, vol. 1, pp. 115-126.

_____, _____, _____, and _____. 1984. "Choosing Between the Sample-Selection Model and the Multi-part Model," Journal of Business and Economic Statistics, vol. 2, pp. 283-289.

Dudley, L. and C. Montmarquette. 1976. "A Model of the Supply of Bilateral Foreign Aid," American Economic Review, vol. 66, pp. 132-142.

Hausman, J. A. 1978. "Specification Tests in Econometrics," Econometrica, vol. 46, pp. 1251-1271.

_____, B. Hall and Z. Griliches. 1984. "Econometric Methods for Count Data with an Application to the Patents-R&D Relationship," Econometrica, vol. 52, pp. 909-938.

_____, B. Ostro and D. Wise. 1984. "Air Pollution and Lost Work," NBER working paper 1263, January.

Hay, J. W. and R. J. Olsen. 1984. "Let Them Eat Cake: A Note on Comparing Alternative Models of the Demand for Medical Care," Journal of Business and Economic Statistics, vol. 2, pp. 279-282.

Heckman, J. 1976. "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models," Annals of Economic and Social Measurement, vol. 5, pp. 475-492.

_____. 1979. "Sample Selection Bias as a Specification Error," Econometrica, vol. 47, pp. 153-161.

Hurd, M. 1979. "Estimation in Truncated Samples When There is Heteroscedasticity," Journal of Econometrics, vol. 11, pp. 247-258.

Johnson, N. L. and S. Kotz. 1969. Distributions in Statistics: Discrete Distributions (New York: Wiley).

_____ and _____. 1970. Distributions in Statistics: Continuous Univariate Distributions - I (New York: Wiley).

Kendall, M. G. and A. Stuart. 1973. Advanced Theory of Statistics, vol. 3 (London: Griffin).

Killingsworth, M. R. 1983. Labor Supply (Cambridge: Cambridge University Press).

Lin, T.-F. and P. Schmidt. 1984. "A Test of the Tobit specification

Against an Alternative Suggested by Cragg," Review of Economics and Statistics, vol. 66, pp. 174-177.

Maddala, G. S. 1977. Econometrics (New York: McGraw-Hill).

_____. 1983. Limited-Dependent and Qualitative Variables in Econometrics (Cambridge: Cambridge University Press).

Manski, C. F. and D. McFadden. 1981. Structural Analysis of Discrete Data with Econometric Applications (Cambridge, Mass: MIT Press).

Morey, E. R. 1981. "The Demand for Site-Specific Recreational Activities: A Characteristics Approach," Journal of Environmental Economics and Management, vol. 8, pp. 345-371.

Nelson, F. D. 1981. "A Test for Misspecification in the Censored-Normal Model," Econometrica, vol. 49, pp. 1317-1329.

Olsen, R. 1980. "Approximating a Truncated Normal Regression with the Method of Moments," Econometrica, vol. 48, pp. 1099-1106.

Ostro, B. 1983. "The Effects of Air Pollution on Work Loss and Morbidity," Journal of Environmental Economics and Management, vol. 10, pp. 371-382.

Pearson, K. and A. Lee. 1908. "Generalized Probable Error in Multiple Normal Correlations," Biometrika, vol. 6, pp. 59-68.

Pitt, M. 1983. "Food Preferences and Nutrition in Rural Bangladesh," Review of Economics and Statistics, vol. 65, pp. 105-114.

Portney, P. R. and J. Mullahy. 1985. "Urban Air Quality and Acute Respiratory Illness," Journal of Urban Economics, forthcoming.

Rao, C. R. 1965. Linear Statistical Inference and Its Applications, (New York: Wiley).

Rosenzweig, M. R. and K. I. Wolpin. 1984. "Migration Selectivity and the Effects of Public Programs," University of Minnesota, Economic

Development Center, Bulletin 84-5.

Rosett, R. N. and F. D. Nelson. 1975. "Estimation of a Two-Limit Probit Regression Model," Econometrica, vol. 43, pp. 141-146.

Royall, R. 1984. "Robust Inference Using Maximum Likelihood Estimators," Johns Hopkins University, Department of Biostatistics Working Paper.

Schmidt, P. 1976. Econometrics (New York: Marcel Dekker).

Smith, M. and G. Maddala. 1983. "Multiple Model Testing for Non-Nested Heteroscedastic Censored Regression Models," Journal of Econometrics, vol. 21, pp. 71-81.

Stapleton, D. and D. Young. 1984. "Censored Normal Regression with Measurement Error on the Dependent Variable," Econometrica, vol. 52, pp. 737-760.

Stewart, M. B. 1983. On Least Squares Estimation when the Dependent Variable is Grouped," Review of Economics Studies, vol. 50, pp. 737-753.

Tobin, J. 1957. "Estimation of Relationships for Limited Dependent Variables," Econometrica, vol. 26, pp. 24-36.

Wales, T. and A. Woodland. 1983. "Estimation of Consumer Demand Systems with Binding Non-Negativitay Constraints," Journal of Econometrics, vol. 21, pp. 263-285.

White, H. 1982. "Maximum Likelihood Estimation of Misspecified Models," Econometrica, vol. 50, pp. 1-25.

White, H. 1983. "Corrigendum," Econometrica, vol. 51, p. 513.

Chapter 3

AIR POLLUTION MONITORS AND INDIVIDUAL EXPOSURES

The models estimated in Volume I typically utilized as measures of an individual's exposure the pollutant-specific readings from the monitor nearest the centroid of the respondent's census tract for which the data were available. In most cases, screening criteria were established so that it was necessary both for a monitor to have recorded at least some minimal number of hourly readings during the two-week period and for the monitor to be located not further than some prescribed distance (20 miles; 10 miles) from the residents' census tract centroids.

It is possible that the nearest-monitor readings we utilized are not representative of the pollution "profile" of the metropolitan area in which each respondent lives. If some average of the readings from a number of nearby monitors better characterizes the ambient concentrations facing the individuals in question, then the consistency of results obtained using nearest-monitor readings must be called into question. (This abstracts, of course, from the larger question of the ability of ambient monitors at all to measure the exposure of individuals.)

The purpose of this very brief chapter and assess the pollution profiles constructed using the nearest-monitor readings versus those that result when the average readings from a number of monitors are used. The extent to which the two constructs are correlated indicates the sensitivity of our results to the use of nearest-monitor readings to characterize exposure.

The procedure is as follows. For each of the six pollutants used in our study--ozone, sulfates, TSP, NO_2, CO, and SO_2--we utilized the data for the 14,441 adults in the main sample and constructed the nearest-monitor measures used in the main study. These were designated as XXNR01, where XX is the specific pollutant (O3,S4,SP,N2,CO,S2). In our study, recall, these measures were subjected to a miles-from-census-tract-centroid cutoff of 5, 10, or, most often, 20 miles; the specific distance will be obvious from the context. (No minimal hours standard is used here.)

For these same individuals, we then constructed two averaged measures for each pollutant. The two measures constructed were the simple average over all the available readings from monitors within 10 and then within 20 miles of the census tract centroids. These measures were designated XXAVGYY, where XX was as defined above and YY was either 10 or 20. Thus, N2AVG20 is the average of all nitrogen dioxide monitors within 20 miles of the census tract centroid.

Then, given these measures, we calculated for each pollutant the correlation between the nearest-monitor reading and the area-average reading at both the 10 and 20 miles cutoff values. We also calculated the number of monitors used to construct the two area-averages.

The results are presented in the tables that follow. In each case, "r" is the simple correlation coefficient between the nearest-monitored reading and the 10 or 20 mile averaged readings.

OZONE

10 mile: r = .965, N = 8,823

Number Monitor Readings in Area-Average:

n	f(n)	n	f(n)
1	3832	6	244
2	2303	7	262
3	985	8	179
4	553	9	112
5	302	10	51

	Mean	Max	Min
03NR01	.0454	.251	0
03AVG10	.0460	.236	0

- -

20 mile: r = .931, N = 11,241

Number Monitor Readings in Area-Average:

n	f(n)	n	f(n)
1	2665	6	679
2	2463	7	646
3	1447	8	550
4	954	9	631
5	847	10	359

	Mean	Max	Min
03NR01	.0450	.251	0
03AVG20	.0461	.225	.003

SULFATES

10 mile: r = .952, N = 5,249

Number Monitor Readings in Area-Average:

n	f(n)	n	f(n)
1	2693	6	134
2	1134	7	109
3	401	8	101
4	308	9	116
5	247	10	6

	Mean	Max	Min
S4NR01	10.528	52.136	0
S4AVG10	10.544	52.136	0

- -

20 mile: r = .912, N = 7,512

Number Monitor Readings in Area-Readings:

n	f(n)	n	f(n)
1	2595	6	526
2	1230	7	559
3	823	8	329
4	614	9	250
5	559	10	27

	Mean	Max	Min
S4NR01	10.590	52.136	0
S4AVG20	10.523	52.136	1.586

TSP

10 mile: r = .859, N = 12,598

Number Monitor Readings in Area-Average:

n	f(n)	n	f(n)
1	1851	6	1039
2	1283	7	975
3	1275	8	1207
4	979	9	1455
5	1012	10	1522

	Mean	Max	Min
SPNR01	70.478	284.004	9.996
SPAVG10	72.021	253.428	15.092

- -

20 mile: r = .818 N = 13.772

Number Monitor Readings in Area-Average:

n	f(n)	n	f(n)
1	822	6	1272
2	895	7	1373
3	702	8	1429
4	797	9	2330
5	1084	10	3068

	Mean	Max	Min
SPNR01	70.128	284.004	9.996
SPAVG20	71.948	272.244	15.092

NO_2

10 mile: r = .951 N = 6,393

Number Monitor Readings in Area-Average:

n	f(n)	n	f(n)
1	3195	6	96
2	1593	7	36
3	775	8	1
4	485	9	0
5	212	10	0

	Mean	Max	Min
N2NR01	117.857	435.316	0
N2AVG10	117.323	435.316	0

- -

20 mile: r = .923 N = 8,452

Number Monitor Readings in Area-Readings:

n	f(n)	n	f(n)
1	3004	6	668
2	1890	7	692
3	747	8	398
4	442	9	168
5	417	10	26

	Mean	Max	Min
N2NR01	112.913	435.316	0
N2AVG20	111.646	375.928	0

CO

10 mile: r = .887 N = 8,921

Number Monitor Readings in Area-Average:

n	f(n)	n	f(n)
1	3638	6	288
2	2087	7	280
3	946	8	235
4	676	9	146
5	510	10	115

	Mean	Max	Min
CONR01	3.806	26.583	0
COAVG10	3.937	26.583	0

20 mile: r = .838 N = 10,939

Number Monitor Readings in Area-Average:

n	f(n)	n	f(n)
1	2344	6	480
2	2536	7	298
3	1130	8	553
4	1215	9	838
5	984	10	561

	Mean	Max	Min
CONR01	3.717	26.583	0
COAVG20	3.808	25.111	0

SO_2

10 mile: r = .857 N = 8,842

Number Monitor Readings in Area-Average:

n	f(n)	n	f(n)
1	2976	6	356
2	1855	7	182
3	1280	8	206
4	966	9	192
5	770	10	59

	Mean	Max	Min
S2NR01	68.591	760.088	0
S2AVG10	69.375	568.988	0

- -

20 mile: r = .819 N = 10,784

Number Monitor Readings in Area-Average:

n	f(n)	n	f(n)
1	2414	6	919
2	1733	7	726
3	1204	8	841
4	1129	9	541
5	1069	10	208

	Mean	Max	Min
S2NR01	66.222	760.088	0
S2AVG20	67.050	568.988	0

The results are quite reassuring about the use of nearest-monitor data to proxy individual exposure. The correlation coefficients between the nearest-monitor reading and the average of all monitors within 10 miles range from .965 for ozone (a highly dispersed pollutant) to .86 for TSP and SO_2 (more localized pollutants). The 20-mile correlations follow similar relationships, but are of course somewhat lower than the 10-mile correlations due to the decreased weight of the nearest-monitor reading in calculating the 20-mile averages. What is particularly encouraging is that no correlation coefficient is below 0.8, leading us to suspect that the use of the nearest-monitor reading would be unlikely to impart any systematic biases vis-a-vis use of area-averaged readings. In the following chapter, we make use of the area-averaged readings to test this suspicion.

Chapter 4

URBAN AIR QUALITY AND ACUTE RESPIRATORY ILLNESS

4.1 Introduction

Over the past fifteen years, economists interested in the benefits of air pollution control have concerned themselves with more than just the appropriate valuation of health gains and losses. In addition, some have explored in epidemiological analyses the actual physical relationships between air pollution and health itself using statistical techniques common in the social and natural sciences. Most of these studies have used aggregate data at the city or SMSA level to test for the effects of prolonged exposures to air pollution on the mortality rates across the units of observation. The studies of Lave and Seskin [8], Crocker, _et al._ [2], Mendelsohn and Orcutt [12], Chappie and Lave [1], and Lipfert [10] are among the best examples.

Relatively less attention has been given in this literature to the relationship between air pollution and sickness (or morbidity). This is unfortunate because morbidity is observed much more frequently than mortality and may be of greater economic significance than premature death. When researchers have examined possible links between air pollution and morbidity, they have generally been forced through lack of data to do so in the absence of information about individuals' socioeconomic and other characteristics—even though these characteristics may have an important effect on health status.

Volume I presents our recently completed comprehensive investigation of the effects of ozone (ground-level rather than stratospheric) and other air pollutants on individuals' acute and chronic health status. Unlike many previous studies, this work is based on a large and relatively detailed individual data base, allowing controls for certain important socioeconomic and demographic characteristics in addition to the meteorological measures sometimes included in earlier studies using either aggregate or less detailed individual data. This chapter presents some of the major findings concerning the effects of urban air quality on acute respiratory disease using an estimation technique not employed in Volume I. Chapter 7 reports some new findings on air pollution and chronic illness.

Of particular concern here is the sensitivity of the findings to the measures of air quality used. As suggested above, most previous analyses of the health effects associated with air pollution have characterized individual exposures using some measure of air quality averaged over most or all of the monitors in the urban areas where the individuals live. However, many persons may get most or all of their ambient exposure proximate to the monitors nearest their homes. As part of our larger study in Volume I, therefore, each individual in the sample was matched to the nearest ten air pollution monitors for each of eight different air pollutants so as to use close-to-home pollution readings to characterize exposure. Because this was very resource-intensive, it is important to illustrate the difference such an approach may make when estimating dose-response relationships. Additional sensitivity analyses in this chapter explore interactive effects as well as possible thresholds and non-linearities in the relationship between air pollution and acute health status.

In Section II we briefly describe the data used in our analysis and the independent variables we include. In Section III, we discuss the estimating techniques used to explore possible links between air pollution and acute respiratory disease. In Section IV we present our empirical findings and in Section V we draw some cautious inferences from them for applied welfare calculations.

4-2. Framework for the Analysis

The individual data underlying both our larger study as well as the present chapter come from the 1979 Health Interview Survey (HIS)--a nationwide sample of approximately 110,000 individuals conducted during the course of each year by the National Center for Health Statistics. All acute illness experienced during the two-week period prior to the date of each interview was to be reported by each respondent or the family member responding for him or her. Manifestations of these illnesses were classified in three types--bed disability days (the most serious of the three categories), work or school loss days, and what might best be thought of as minor restricted activity days. The latter are days on which the respondent was neither bed-ridden nor forced to miss work or school but did suffer from an acute impairment sufficient to cause him or her to restrict activity in some noticeable way. The dependent variable in the subsequent analysis is total restricted activity days--the total number of days during the two-week period on which any of these three types of acute illness occurred. Finally, all acute (and chronic) health information elicited in the survey was coded by cause, using the International Classification of Disease. Attention is limited in this chapter to total restricted activity days due to respiratory disease since this is the type of acute impairment most likely to result from exposure to air pollution.

The socioeconomic data elicited from each respondent in the Health Interview Survey includes, among many other individual and household-specific characteristics, information on age, race, sex, income, and education. In addition, several supplements to the 1979 survey made it particularly useful for epidemiological purposes. Specifically, the 1979 HIS contained a supplemental questionnaire asked of one-third of all the adults interviewed (26,271 of a total of 79,743 adults) which provided detailed data on lifetime smoking history, including the tar and nicotine content of the brands most commonly smoked. Smoking data are obviously of great importance if one is interested in exploring the determinants of respiratory and other types of disease. The 1979 HIS also included a supplement (again to one-third of all adults surveyed) designed to provide detailed information on residential histories. This is not important for our present purposes but will play a major role in our analysis in Chapter 7 of the determinants of chronic respiratory and other types of disease.

All air pollution data come from the Environmental Protection Agency's SAROAD system. For our analysis of the relationship between air pollution and acute morbidity in the larger study, all air quality data were specific to the two-week recall period for which individual health data were available. This is also the case here, save for sensitivity analyses conducted using annual average data as a proxy for air quality during the two-week period. As indicated above, most of the analysis below characterizes individuals' exposures to air pollution using data from the air pollution monitors nearest their residences. No individuals are included in the final sample if the nearest monitor for any pollutant is more than ten miles away. The average distance to the nearest monitor is slightly more than four miles. In addition to the air pollution data, meteorological data were added from the monitoring network of the National Oceanic and Atmospheric Administration. Included are observations on temperature and precipitation during the two-week recall period.

The overall sample from which the subsample used here is drawn consists of 14,441 individuals aged seventeen and above for whom both smoking data and at least some air pollution data were available. The models estimated below are based on a smaller subsample, however, since complete data are required for each of the air pollutants and other independent variables.

The analysis of acute respiratory disease includes air pollution data during the two-week recall period for ozone, a gaseous pollutant that is the primary constituent of smog, and well as sulfates, perhaps the most harmful of the airborne particles. It is worth noting that the computer algorithm used to match individuals to the ten nearest ozone and sulfate monitors could only be used for monitors within SMSAs. Thus, the estimation sample consists entirely of city and suburban residents from around the United States. Table 4-1 lists the independent variables used in the analysis of acute respiratory disease and their sample means.

Table 4-1. Variable Definitions and Sample Means

Variable Name	Description	Sample Mean
OZNEAR	Average daily maximum one-hour ozone reading during two week recall period at monitor nearest the centroid of respondent's census tract of residence (in parts per million)	0.042
S4NEAR	Average 24-hour sulfate concentration during two weeks at nearest monitor (see above) (in micrograms per cubic meter)	10.876
OZAV10	Average daily maximum one-hour ozone reading during two weeks averaged over all monitors within a ten mile radius of respondent's census tract centroid	0.043
S4AV10	Average 24-hour sulfate concentration during two weeks averaged as in OZAV10	10.890
OZAV20	Same as OZAV10 but averaged over all monitors within 20 mile radius	0.044
S4AV20	Same as S4AV10 but averaged over all monitors within 20 mile radius	10.700
OZANNR	Average daily maximum one-hour ozone concentration over entire calendar year 1979 as measured at the nearest monitor	0.042
S4ANNR	Average 24-hour sulfate concentration over calendar year 1979 as measured at the nearest monitor	10.752
OZAN10	Same as OZANNR but averaged over all monitors within ten mile radius	0.043
S4AN10	Same as S4ANNR but averaged over all monitors within 10 miles	10.709
OZAN20	Same as OZAN10 but averaged over all monitors within 20 miles	0.044

73<

Table 4-1 (cont'd). Variable Definitions and Sample Means

Variable Name	Description	Sample Mean
S4AN20	Same as S4AN10 but averaged over all monitors within 20 miles	10.588
WHITE	Dummy variable: 1 if white, 0 otherwise	0.852
MALE	Dummy variable: 1 if male, 0 if female	0.436
INCOME	Annual household income in dollars	17,152
AGE	Age in years	42.80
CIGS	Number of cigarettes smoked per day	7.58
FORMER	Dummy variable: 1 if respondent formerly smoked regularly but does not presently, 0 if not	0.20
SCHLYR	Years of education completed	11.73
CHRONIC	Dummy variable: 1 if respondent has any limitation in activity due to chronic illness, 0 otherwise	0.17
MAXTMP	Average daily maximum temperature during two-week period	64.02
RAIN	Average daily rainfall during two-week period	0.12
RRAD	Number of respiratory-related restricted activity days during two-week recall period	0.162

4.3 Model Specification

For reasons of economy and computational simplicity, most of the models in Volume I were estimated using ordinary least squares and logit techniques (where the dependent variable was, respectively, either the number of days of a particular kind of impairment during the two-week recall period or a dichotomous indicator of an individual having at least one day of that kind of impairment during the period). As Chapter 2 points out, however, estimation techniques like OLS are not ideally suited to the nature of our measures of acute health status, however. Recall that that measure is the number of respiratory-related restricted activity days during the two-week recall period (RRADs). Clearly this measure is bounded by zero and fourteen and because of survey protocol can assume only integer values in $\{0,1,2,\ldots,14\}$. The frequency distribution of RRADs for the sample of 3,347 adults is presented in Table 4-2. Because of the small number of observations at the upper (14 day) limit, the implications of this upper bound for estimation strategy are ignored in the following analysis; we concentrate instead on the complications arising from the overwhelmingly large number of individuals reporting zero RRADs.

A standard approach in such circumstances is to use the Tobit or censored normal estimator where one observes T independent observations on y_t which are the realizations of random variables Y_t^* subject to the censoring rule $y_t=\max(0,y_t^*)$, $Y_t^*-N(X_t\beta,\sigma^2)$. However, parameter estimates obtained using the Tobit model are generally inconsistent when the underlying data are not distributed as censored normal with

Table 4-2. RRAD Frequency Distribution

RRAD	# OBS	%
0	3227	96.42
1	25	0.75
2	28	0.84
3	23	0.69
4	9	0.27
5	7	0.21
6	2	0.06
7	3	0.09
8	3	0.09
10	3	0.09
11	1	0.03
12	1	0.03
14	15	0.45

independent, identically distributed errors. Estimating a Tobit model of RRADs using the two-week average pollution data from the nearest monitor and the other independent variables in Table 4-1 above, some tests for its appropriateness were conducted and strong evidence of misspecification was found. While this might be attributable to omitted variables or other factors unrelated to departures from the usual assumptions about the error distribution in the Tobit model, a different statistical approach is utilized here.

In modeling event counts (non-negative integer data) over a time interval (t,t+dt), the Poisson distribution is commonly used. Here, discrete random variates Y_t follow the probability law:

$$(1) \qquad \Pr(Y_t = y) = \begin{cases} \exp(-\lambda_t)\lambda_t^y/y!, & y = 0,1,2,\ldots \\ 0, & \text{else} \end{cases}$$

with $E(Y_t) = \mathrm{Var}(Y_t) = \lambda_t$. Given the nonnegative integer nature of the RRAD measure, such a probability law has obvious appeal for estimation. Analogous to the normal distribution where for econometric work one typically specifies $E(Y_t) = \mu_t = X_t\beta$, the parameter of the Poisson distribution can be reparameterized to admit the influence of covariates. Since for all t, $\lambda_t > 0$, a straightforward approach is to assume $\lambda_t = \exp(X_t\beta)$ and to estimate β by maximum likelihood (see Hausman, Hall, Griliches [5], Hausman, Ostro, Wise [4]). This is the approach adopted here for modeling the RRAD outcomes.

A drawback of the Poisson specification is the restriction that $E(Y_t) = Var(Y_t)$. Should this restriction not characterize the data, the maximum likelihood estimates of the covariance matrix of $\hat{\beta}$ will be inconsistent and asymptotic t-tests based thereon would be misleading. Hausman, Ostro, and Wise circumvent this restriction by allowing for an overdispersion parameter. We take a different approach here, using an estimator of the covariance matrix that is more robust against departures from the restriction that the mean be equal to the variance. Details of this procedure are presented in the appendix.

Given the assumptions on the parameterization of the λ_t, the log-likelihood function to be maximized is:

$$(2) \qquad \ell = \sum_t -\exp(X_t\beta) + y_t X_t \beta + c,$$

where X_t is the vector of independent variables as described in Table 1, y_t is the observed RRAD count for individual t, and c does not depend on β. The ML estimate of β satisfies:

$$(3) \qquad \partial\ell/\partial\beta = \sum_t (-\exp(X_t\beta) + y_t)X_t' = 0.$$

4.4. Empirical Results

Table 4-3 presents the results of our basic model and the variants designed to test the sensitivity of the results to assumptions about individual exposures to ambient air pollution. In specifications (3.1) - (3.3) each individual's count of respiratory restricted activity days is hypothesized to be related to ambient air quality during the individual's two-week recall period. In (3.1) exposures are proxied by readings from the one ozone and one sulfate monitor nearest each individual's residence; in specifications (3.2) and (3.3), readings are averaged, respectively, over all monitors within 10 and 20 miles of each respondent's residence. Specifications (3.4) - (3.6) use annual 1979 average air pollution readings as a proxy for air pollution exposure during each recall period. As in (3.1) - (3.3), equation (3.4) uses the annual average at the nearest monitor to proxy individual exposure while (3.5) and (3.6) use the average of the annual averages of all monitors within 10 and 20 miles respectively.

Table 4-3 indicates that of the non-pollution variables, race, income and temperature are related consistently across models to RRADs in a statistically significant way—with whites, those with lower incomes, and those exposed to colder temperatures all experiencing relatively more acute respiratory illness during the two-week recall period. Because those reporting the presence of a chronic illness would be expected to experience more restrictions in activity during any two-week period, a dummy variable identifying such individuals was included. As expected, this dummy variable was positively and significantly related to the number of RRADs. Finally, while always of the expected sign, the number of cigarettes smoked

Table 4-3. Model Estimates: Sensitivity to Air Pollution Measurement
(Dependent variable is RRADs during two-week recall period)

Independent Variable	Model 3.1	 3.2	 3.3	 3.4	 3.5	 3.6
OZNEAR	6.883					
	(1.97)					
OZAV10		6.614				
		(1.91)				
OZAV20			9.324			
			(2.41)			
OZANNR				17.603		
				(3.18)		
OZAN10					19.449	
					(2.88)	
OZAN20						17.473
						(2.12)
S4NEAR	-0.005					
	(0.22)					
S4AV10		-0.0210				
		(0.67)				
S4AV20			-0.046			
			(1.4)			
S4ANNR				-0.0175		
				(0.41)		
S4AN10					-0.0558	
					(1.34)	
S4AN20						-0.0765
						(1.87)
WHITE	1.261	1.258	1.249	1.165	1.163	1.188
	(2.87)	(2.86)	(2.85)	(2.65)	(2.65)	(2.72)

Table 4-3 (cont'd.) Model Estimates: Sensitivity to Air Pollution Measurement (Dependent variable is RRADs during two-week recall period)

Independent Variable	Model 3.1	 3.2	 3.3	 3.4	 3.5	 3.6
MALE	-0.054 (0.19)	-0.058 (0.21)	-0.064 (0.23)	-0.062 (0.22)	-0.065 (0.23)	-0.055 (0.20)
INCOME	-0.000035 (2.31)	-0.000035 (2.30)	-0.000035 (2.28)	-0.000035 (2.27)	-0.000034 (2.22)	-0.000033 (2.19)
AGE	0.00031 (0.05)	0.00050 (0.08)	0.00086 (0.14)	0.00076 (0.13)	0.0013 (0.23)	0.0013 (0.21)
CIGS	0.015 (1.53)	0.015 (1.56)	0.016 (1.62)	0.016 (1.71)	0.016 (1.70)	0.016 (1.72)
FORMER	0.312 (0.89)	0.319 (0.91)	0.323 (0.92)	0.340 (0.98)	0.344 (0.99)	0.328 (0.94)
SCHLYR	0.0067 (0.17)	0.0066 (0.17)	0.0067 (0.17)	.000062 (0.02)	0.0035 (0.09)	0.0010 (0.03)
CHRONIC	0.776 (2.45)	0.769 (2.42)	0.760 (2.39)	0.707 (2.18)	0.071 (2.19)	0.720 (2.24)
MAXTMP	-0.019 (2.45)	-0.018 (2.54)	-0.021 (2.70)	-0.016 (2.18)	-0.017 (2.38)	-0.018 (2.48)
RAIN	1.629 (1.07)	1.735 (1.13)	1.952 (1.28)	1.801 (1.12)	1.992 (1.30)	2.049 (1.34)
INTERCEPT	-2.127 (2.06)	-1.993 (1.92)	-1.780 (1.67)	-2.559 (2.26)	-2.257 (1.93)	-1.950 (1.66)
N	3,347	3,347	3,347	3,347	3,347	3,347
L^*	-741.5	-740.9	-732.0	-710.0	-707.0	-712.0

L^* = Log likelihood

(Asymptotic normal statistics for Ho: β_k =0 in parentheses) per day and the dummy variable indicating that the respondent is a former smoker were not significant at conventional levels, a somewhat surprising finding given the concentration on respiratory disease.

The main focus of our analysis is the relationship between acute respiratory disease (as measured by RRADs) and urban air quality. As Table 4-3 indicates, in only one of the six specifications is the hypothesis of no relationship between ozone and RRADs not rejected at at least the 95% level.[2] This finding is fully consistent with the analysis in Volume I where we used different samples, estimating techniques, and combinations of independent variables—including monitored readings for as many as five separate air pollutants. There positive and significant associations between ozone and RRADS in adults were frequent although not uniform.

The statistical significance of the ozone coefficients is not altered appreciably by using monitored readings averaged over 10 or 20 miles rather than readings at the nearest monitor. This is intuitively plausible since ozone tends to be a diffuse (as opposed to a "hot-spot") pollutant. To the extent they are generalizable, our findings suggest that city or SMSA-wide average readings may be preferable to nearest-monitor readings to characterize individual exposure to ozone in view of the resources required to obtain the latter.

Using air pollution data averaged over the entire year during which the health interview took place—models (3.4) - (3.6)—results in larger estimated coefficients and higher asymptotic t-ratios for the ozone variable than when air quality data contemporaneous to the recall period are used.[3] The importance of this finding should be discounted, we believe. So long as one is concerned with the possible relationships between urban air quality and day-to-day variations in acute morbidity, the correct measure of pollution must be one which is coincident with, or slightly precedes, the period during which health status is being observed. To illustrate, consider an individual interviewed for the HIS on January 15, 1979. Clearly, using 1979 annual average air pollution readings for ozone and sulfates to help explain RRADs between January 1-14 brings into play 50 weeks of data which could have no effect whatsoever on health during the recall period. For this reason, the use of contemporaneous (or "real time") air pollution data should be considered the conceptually correct approach when analyzing acute respiratory disease.

Based on the findings in Table 4-3 we cannot reject the hypothesis of no relationship between ambient sulfate concentrations and RRADS during the two-week recall period. It should be noted, however, that sulfates and other particulates are generally monitored only every six days. Thus, any two-week period will contain at most three 24-hour sulfate measurements and this may affect the findings. (Ozone, on the other hand, is monitored continuously and is measured in specifications (3.1) - (3.6) by the average daily maximum one-hour reading--measured during the recall period or annually depending on the equation.) Note also that the coefficient on sulfates is more sensitive to the choice of exposure proxy. This is because concentrations of sulfates and other particulates exhibit greater variation within an area than does ozone.[4] (It should be noted here that the sample correlation between OZNEAR and S4NEAR is 0.108. We conducted tests for possible degradation of parameter estimates due to collinearity but found no evidence thereof.)

Prior clinical and epidemiological analyses suggest the possible importance of interactive or synergistic effects of certain air pollutants (see Hazucha and Bates [7] and Graves and Krumm [3], for instance). Accordingly, the existence of such an effect between ozone and sulfates (OZXS4) is tested. The results are presented in specification (4.1) in Table 4-4, and do not support the hypothesis that such effects are important. In (4.2) another hypothesized interactive effect is tested, that between ozone and average maximum temperature (OZXTEMP) during the recall period. Again, no evidence of such an effect is found. These results are consistent with the more extensive analysis of interactive effects in Volume I.

So-called threshold effects or other types of non-linearities in the relationship between ozone and RRADs are potentially important and are tested for here. To see whether the relationship with RRADs differs between lower and higher concentrations, the sample was twice divided into two separate regimes, once with the dividing point being 0.05 ppm. Separate coefficients were estimated on the ozone variable in the lower and higher regimes. In this specification ozone is positively and significantly associated with the expected number of RRADs in regimes both above and below 0.05 ppm. A causal inspection of the coefficients in (4.3) could convey the impression that a marginal change in ozone will have a larger impact on RRADs at lower than at higher concentrations. In fact, this is not the case. When the first derivatives of the estimating

Table 4-4. Model Estimates: Alternative Specifications (Dependent variable is RRADs during two-week recall period)

Independent Variable	Model 4.1	 4.2	 4.3	 4.4	 4.5
OZNEAR	7.410 (1.24)	70.659 (1.77)			
OZH05			9.554 (2.71)		
OZL05			22.505 (2.11)		
$(\mathrm{OZNEAR})^2$				1.343 (0.07)	
$(\mathrm{OZNEAR})^{1/2}$					4.926 (2.45)
S4NEAR	−0.003 (0.09)	−0.003 (0.12)	−0.0023 (0.10)	−0.0017 (0.07)	−0.0074 (0.31)
OZXS4	−0.047 (0.09)				
OZXTEMP		−0.874 (1.65)			
WHITE	1.262 (2.87)	1.235 (2.80)	1.259 (2.86)	1.290 (2.93)	1.239 (2.83)
MALE	−0.054 (0.19)	−0.053 (0.19)	−0.049 (0.18)	−0.043 (0.15)	−0.060 (0.21)
INCOME	−0.000035 (2.31)	−0.000036 (2.37)	−0.000036 (2.38)	−0.000035 (2.32)	−0.000036 (2.32)
AGE	0.00031 (0.05)	0.00067 (0.11)	−0.000024 (0.04)	0.00025 (0.04)	0.00034 (0.06)

Table 4-4 Cont'd.) Model Estimates: Alternative Specifications (Dependent variable is RRADs during two-week recall period)

	Model				
Independent Variable	4.1	4.2	4.3	4.4	4.5
CIGS	0.015 (1.53)	0.015 (1.52)	0.015 (1.52)	0.014 (1.49)	0.151 (1.55)
FORMER	0.312 (0.89)	0.318 (0.90)	0.318 (0.90)	0.303 (0.87)	0.319 (0.91)
SCHLYR	0.0067 (0.17)	0.0036 (0.09)	0.0051 (0.13)	0.0063 (0.16)	0.0071 (0.18)
CHRONIC	0.776 (2.44)	0.779 (2.49)	0.773 (2.44)	0.773 (2.43)	0.781 (2.47)
MAXTMP	-0.019 (2.44)	0.0059 (0.32)	-0.019 (2.50)	-0.013 (1.85)	-0.023 (2.92)
RAIN	1.632 (1.07)	1.763 (1.12)	1.626 (1.07)	1.366 (0.87)	1.776 (1.17)
INTERCEPT	-2.152 (1.92)	-3.827 (2.29)	-2.489 (2.26)	-2.225 (2.14)	-2.498 (2.46)
N	3,347	3,347	3,347	3,347	3,347
ℓ	-2049.4	-2031.2	-2039.2	-2054.3	-2043.1

(Asymptotic normal statistics for Ho:β_K =0 in parentheses) equation are evaluated at the appropriate ozone concentration for each of the individuals in the low and high regimes and the resulting values then averaged, the estimated first derivative is nearly twice as large in the high as in the low regime.

Although the Poisson expectation function $E(RRAD_t) = \exp(X_t\beta)$ is non-linear, it does imply that the elasticity of $E(RRAD_t)$ with respect to ozone is linear. To allow for greater flexibility, models (4.4) and (4.5) are estimated using, respectively, the square and the square root of the ozone concentration during the recall period at the nearest monitor. In other words, the specification is:

$$(4) \quad E(RRAD_t) = \exp(Z_t\gamma + \alpha(OZNEAR_t)^{\delta})$$

where Z_t is the vector of independent variables other than ozone and $\delta\epsilon(0.5, 2.0)$. When $\delta=0.5$, $\hat{\alpha}$ is positive and statistically significant; when $\delta=2.0$, $\hat{\alpha}$ is positive but not significant. In fact, note that (4.5) has a higher model likelihood than specification (3.1) which is simply equation (4) with $\delta=1.0$, thus indicating that non-linearities in the ozone specification are important.[5]

4.5. Policy Implications and Conclusions

Ozone is one of six air pollutants for which the Environmental Protection Agency has established maximum permissible ambient concentrations. The controversy surrounding revision of the ozone standard in 1978 (see White [20]), coupled with recent emphasis on cost-benefit analysis in government regulation (see Smith [17]), make it worthwhile to illustrate the changes in acute respiratory health that might be associated with changed ozone levels. We use a subset of the results presented above to make such an illustrative calculation. The discussion here is confined to specifications where ozone is measured by the average daily one-hour maximum during the two weeks at the monitor nearest the respondent's residence.

One way to assess possible pollution-related changes in acute health status is to calculate the elasticity of E(RRAD) with respect to ozone and evaluate the predicted total change in expected RRADs for the individuals in the sample resulting from some hypothetical change in ozone concentrations. Log-differentiating (4), it follows that:

$$(5) \quad (\partial E(RRAD_t)/\partial OZNEAR_t)(OZNEAR_t/E(RRAD_t)) = \delta\alpha(OZNEAR_t)^{\delta}.$$

Note that for $\delta<1$, the curvature of the expectation function (as determined by $\partial^2 E(RRAD_t)/\partial OZNEAR_t^2$) cannot be determined without reference to the data for the t-th observation. It can be seen from (5) that in the nonlinear cases where $\delta=0.5$ or 2.0--as in specifications (4.4) and (4.5)--evaluating the elasticity at the sample mean of $OZNEAR_t$ will yield a different estimate than that given by evaluating (5) for all t and then averaging the elasticities. The results of both approaches are presented in the top panel of Table 4-5.

The upper panel of the table indicates that the estimated elasticities are quite sensitive to the value of δ. When δ assumes the values 0.5 and

1.0, the resulting elasticities are of the same order of magnitude, with the former roughly twice the latter. However, when $\delta=2.0$, the estimated elasticity is almost two orders of magnitude smaller than the others. Note that these results hold irrespective of the method of elasticity calculation.

In the lower panel of Table 4-5 are presented the elasticities for the model (4.3) in which ozone was permitted to have different coefficients in low and high regimes. Recall that in this case $\delta=1.0$, so that within each regime the ozone elasticities are linear in ozone. Therefore, both methods used above to calculate elasticities will yield the same result. However, there are two relevant elasticity measures, one prevailing for observations with ozone measures below the split and one for those above. Because of the second-derivative properties noted above, reference to the parameter estimates alone is insufficient to compare low- and high-regime elasticities. In fact, it happens that the elasticity estimates for the low-ozone and high-ozone regimes are virtually identical, 0.65 and 0.66, respectively.

Table 4-5. Elasticity Estimates for Alternative Specifications

Whole Sample

δ	Evaluated at Mean of $OZNEAR_t$	Mean of Individual Elasticities
0.5	0.506	0.485
1.0	0.290	0.290
2.0	0.0048	0.006

Split Sample (w/ = 1.0)

Split =	
0.05 ppm	
low regime	0.645
high regime	0.655
0.075 ppm	
low regime	1.061
high regime	0.209

Table 4-6. Estimated Changes in RRADs Due to 10 percent Reduction in Ozone Concentration

δ=	Average Individual Reduction each two weeks $(S_1-S_2)/n$	Annual Decrease in RRADs: Urban Adult Population*
0.5	-.00776	22.19×10^6
1.0	-.00442	12.64×10^6
2.0	-.000083	0.24×10^6

*Calculated by multiplying the two-week individual change in column 2 by 26 to convert to annual changes and then by 100 million--the urban adult population of the United States.

The elasticity estimates can be used to estimate one type of health improvement that might accompany reduced ozone concentrations. Using the estimates $\hat{\beta} = (\hat{\gamma}, \hat{\alpha})$ from the specifications (3.1), (4.4), and (4.5), (5) is evaluated at (Z_t, OZNEAR_t) for all t in the estimation sample and the sum $S1 = \sum_t \exp(Z_t\hat{\gamma} + \hat{\alpha}(\text{OZNEAR}_t)^{\delta})$ is calculated for each of the three alternative specifications. This yields an estimate of the prevailing count of RRADs in the sample of 3,347 adults given prevailing levels of the independent variables including ozone. To evaluate the effect of a change, we first assume that some hypothetical policy measure reduces by 10 percent the two-week average daily maximum ozone concentration, OZNEAR_t, faced by each individual and then calculate the sum $S2 = \sum_t \exp(Z_t\hat{\gamma} + \hat{\alpha}(.9*\text{OZNEAR}_t)^{\delta})$.

For each of the three specifications, the average (S1-S2)/3347 is calculated, thus giving an estimate of a typical individual's change in two-week RRADs given a ten percent decrease in ozone concentrations. Assuming an adult SMSA population of 110 million, and extrapolating the two-week decrease in RRADs to an annual figure, we obtain for each specification an estimate of the total annual decrease in respiratory-related restricted activity days associated with a hypothetical ten percent ozone reduction. The results are presented in Table 4-6.

It is here that the implications of the different specifications can most forcefully be seen. At the two extremes are the δ=0.5 and δ=2.0 formulations of the model. In the former case, the ten percent reduction evokes a total annual change of more than 22 million RRADs while in the latter case the change is less than a quarter million RRADs.

The final step in benefit estimation involves the assignment of dollar values to these hypothetical improvements in health. Valuing reduced RRADs is not easy, particularly since that measure embodies a range of impairments from minor restrictions in activity to bed disability days. However, based on separate analysis of adults' work loss and bed disability in Volume I--wherein we found no significant associations with ambient ozone concentrations and the more severe types of restrictions -- we presume that the effects predicted in Table 4-6 are minor restrictions in activity.

Ideally, these minor RRADs should be valued using changes in individuals' expenditure functions which reflect both labor-leisure tradeoffs as well as the possibility of defending against pollution-related illness (see Harrington and Portney [7], for instance). In practice, alternative approaches are typically required. Using contingent valuation methods, for example, Loehman *et. al.* [11] recently elicited individuals' reported willingness to pay to avoid one day of various kinds of respiratory impairments. The values ranged from $2.31 for a day of minor coughing and sneezing to about $11.00 to avoid a day of severe shortness of breath. Since the latter impairment is likely to be associated with a work loss and/or bed disability day, the former value is probably more appropriate for a minor RRAD. Because of the many uncertainties in arriving at such estimates, however, we assume that a minor RRAD could be valued at as much as $20. If each of a predicted 22 million fewer RRADs are valued at $20, annual benefits to the adult urban population of the U.S. would be $0.44 billion. If RRADs were as few as 250,000 (as predicted in the third row of Table 4-6), and each was valued at $2.31, the corresponding total would be but $0.58 million.

It is important to note that reduced ozone concentrations may result in other beneficial effects besides possible reductions in acute respiratory illness. These include improved visibility, reduced damages to forests, ornamental plantings, and agricultural output,[6] as well as other welfare-enhancing changes. All these would have to be considered (and valued, where appropriate) in any comparison of the costs and benefits of ozone control.

Even when attention is confined to acute respiratory illness, however, the uncertainties in estimating benefits are substantial. Both here and in Volume I, predicted changes in RRADs proved somewhat sensitive to the choice and measurement of independent variables and, in Volume I at least, the size of the sample over which the parameters were estimated. Even when these are held constant, Table 4-6 demonstrates that predicted changes in RRADs are also sensitive to the assumed form of the exposure-response function (by two orders of magnitude). Moreover, this difference is based on a comparison of point estimates without regard to confidence intervals constructed about them. These uncertainties, coupled with sometimes conflicting findings from other epidemiological or clinical studies, must make one cautious in using studies like this in policymaking.

APPENDIX

As described in Section 4.3, the log-likelihood function of the RRAD models can be written as

$$\text{(A1)} \qquad \ell = \sum_t -\exp(X_t\beta) + y_tX_t\beta + c,$$

where $\exp(X_t\beta) = \lambda_t$. It is easy to show that ℓ is concave in β so long as its inverse Hessian exists. As mentioned in Section 4.3, the maximum likelihood estimates of β obtained by maximizing (A1) are consistent, but the estimate of the covariance matrix of $\hat{\beta}_{ML}$ using minus the inverse of the Hessian evaluated at $\hat{\beta}_{ML}$ will tend to be inconsistent if the data are not in fact generated by the specified Poisson distribution.

This is easily seen as follows. Note that the model can be equivalently cast as a nonlinear least squares regression, the t-th observation being

$$\text{(A2)} \qquad \begin{aligned} y_t &= E(Y_t) + u_t \\ &= \exp(X_t\beta) + u_t \end{aligned}$$

with $E(u_t) = 0$. Clearly, $\text{Var}(u_t) = \text{Var}(Y_t) = \exp(X_t\beta)$ so that the u_t are heteroscedastic. If nonlinear weighted least squares is used with the weights $\exp(-X_t\beta)$ formed using consistent estimates of β, and if the data are in fact Poisson-distributed as maintained, the maximum likelihood consistent estimates of β and $\text{Cov}(\hat{\beta})$ will obtain. (The consistency of $\hat{\beta}_{ML}$ for β does not depend on the weighting scheme.) However, if the data are not Poisson-distributed, the estimate of $\mathrm{Cov}(\hat{\beta})$ obtained in this manner will be inconsistent and t-tests based thereon will be misleading. The case is fully analogous to the estimation of the heteroscedastic linear model which yields inconsistent covariance estimates (and, therefore, t-statistics) if the heteroscedastic nature of the error structure is either ignored or incorrectly specified.

White [18] and Royall [16] have demonstrated a method whereby estimates of $\mathrm{Cov}(\hat{\beta})$ robust against misspecification of the underlying distribution of the data can be obtained when $[-\partial^2\ell/\partial\beta\partial\beta']^{-1}$ evaluated at $\hat{\beta}_{ML}$ fails to yield a consistent estimate of $\mathrm{Cov}(\hat{\beta})$. Denoting $I(\beta)$ as $[-\partial^2\ell/\partial\beta\partial\beta']$, their suggestion is to estimate $\mathrm{Cov}(\beta)$ as

$$\text{(A3)} \qquad I(\beta)^{-1}\Big[\sum_t(\partial\ell_t/\partial\beta)(\partial\ell_t/\partial\beta)'\Big]I(\beta)^{-1}$$

where ℓ_t is the t-th observation's contribution to the log-likelihood function and where all relevant evaluations in (A3) are at $\hat{\beta}_{ML}$. This is the method used in constructing the confidence intervals for the parameter estimates of Section 4-4. In these cases, the standard errors of the parameter estimates obtained using $I(\hat{\beta})^{-1}$ as the estimate of $\mathrm{Cov}(\hat{\beta})$ are found to be about two to three times smaller than those obtained using this alternative method. As noted by White [19], the alternative approach (i.e. using (A3)) will typically lead to conservative inferences (i.e. "too large" estimates of $\mathrm{Cov}(\hat{\beta})$) in instances where X_t is nonstochastic and varies across t, as is the case here.

Footnotes

[1] Specifically, the Tobit specification error tests of Nelson [13] and Lin and Schmidt [9] were used. Nelson's is a Hausman test while that of Lin and Schmidt is a Lagrange multiplier test. Under the null hypothesis of no misspecification, both test statistics are distributed asymptotically central $\chi^2_{(k)}$, where k is the dimensionality of β. For the specification described above, both statistics indicate rejection of the no misspecification hypothesis at better than the 98% level.

[2] These confidence intervals are constructed using the approach discussed in the appendix, which should give conservative asymptotic t-statistics. Confidence intervals based on minus the inverse Hessian of the Poisson log-likelihood function, on the other hand, are much tighter, but are almost certainly misleading (inconsistent), given the data used. These results are available from the authors on request.

[3] The substantial discrepancy between the magnitudes of the estimates of the two-week and annual ozone coefficients results, loosely speaking, from the fact that--while the sample means of the two measures are virtually identical--the sample variances of the two-week measures are much larger than those of the annual counterparts in conjunction with the fact that the expectation $E(RRAD_t)$ is the convex function $\exp(X_t\beta)$.

[4] For comparison's sake, model (3.1) was also cast as a geometric distribution. Here, $Pr(Y_t=y) = P^y/(1+P)^{y+1}$ for $y=0,1,2,\ldots$, $E(Y_t) = P$, $var(Y_t) = P(1+P)$, and for purposes of econometric estimation $E(Y_t) = \exp(X_t\beta)$ and $Var(Y_t) = \exp(2X_t\beta) + \exp(X_t\beta)$ are specified. As expected,

the estimated variances of $\hat{\beta}$ were somewhat larger than those obtained using the uncorrected variance version of the Poisson specification while the estimates themselves were quite similar. However, like the Poisson, the maximum likelihood variance estimates based on minus the inverse of the Hessian evaluated at $\hat{\beta}_{ML}$ are not generally consistent if the data are not distributed according to the postulated geometric distribution. Thus, while larger than the estimated variances of the uncorrected Poisson specification, the ML estimates of the geometric parameter variances were still substantially smaller than those obtained using the alternative approach.

[5] δ is, of course, a parameter to be estimated rather than a given constant. The ML algorithm used to obtain the Poisson parameter estimates, however, did not permit estimation of such additional nonlinearities.

[6] Kopp, Raymond, William Vaughan, Michael Hazilla and Richard Carson, "Implications of Environmental Policy for U.S. Agriculture: The Case of Ambient Ozone Standards," Resources for the Future working paper, January 5, 1984.

References

[1] Chappie, Michael and Lester Lave, "The Health Effects of Air Pollution: A Reanalysis," *J. Urban Econ.*, vol. 12 (1982) pp. 346-76.

[2] Crocker, Thomas, *et. al.*, "Methods Development for Assessing Air Pollution Control Benefits," Vol. 1, EPA Document EPA-600/5-79-001a (1979).

[3] Graves, Philip and Ronald Krumm, "Morbidity and Pollution: Model Specification Analysis for Time-Series Data on Hospital Admissions," *J. Environ. Econ. Manage.*, vol. 9 (1982) pp. 311-327.

[4] Hausman, Jerry, Bart Ostro, and David Wise, "Air Pollution and Lost Work," NBER working paper no. 1263, January 1984.

[5] Hausman, Jerry, Bronwyn Hall, and Zvi Griliches, "Econometric Models for Count Data with an Application to the Patents-R&D Relationship," *Econometrica*, vol. 52 (1984) pp. 909-938.

[6] Harrington, Winston and Paul R. Portney, "Valuing the Benefits of Health and Safety Regulation in the Presence of Defensive Expenditures," RFF Quality of the Environment working paper no. QE84-09, September 1984.

[7] Hazucha, Michael and David Bates, "Combined Effects of Ozone and Sulfur Dioxide on Human Pulmonary Function," *Nature*, vol. 257 (1975) pp. 50-51.

[8] Lave, Lester and Eugene Seskin, *Air Pollution and Human Health* (Baltimore, Md.: Johns Hopkins University Press, 1977).

[9] Lin, Tsai-Fen and Peter Schmidt, "A Test of the Tobit Specification Against an Alternative Suggested by Cragg," *Review of Economics and Statistics*, vol. 66 (1984) pp. 174-177.

[10] Lipfert, Frederick, "Air Pollution and Mortality: Specification Searches Using SMSA-Based Data," *J. Environ. Econ. Manage.*, vol. 11 (1984) pp. 208-243.

[11] Loehman, Edna *et. al*, "Distributional Analysis of Regional Benefits and Costs of Air Quality Control," *J. Environ. Econ. Manage.*, vol. 6 (1979) pp. 222-243.

[12] Mendelsohn, Robert and Guy Orcutt, "An Empirical Analysis of Air Pollution Dose Response Curves," *J. Environ. Econ. Manage.*, vol. 6 (1979) pp. 85-106.

[13] Nelson, Forrest, "A Test for Misspecification in the Censored Normal Model," *Econometrica*, vol. 49 (1981) pp. 1317-1330.

[14] Ostro, Bart, "The Effects of Air Pollution on Work Loss and Morbidity," *J. Environ. Econ. Manage.*, vol. 10 (1983) pp. 371-382.

[15] Portney, Paul and John Mullahy, "Ambient Ozone and Human Health: An Epidemiological Analysis," report prepared for Economic Analysis Branch, Office of Air Quality Planning and Standards, USEPA under contract no. 68-02-3583, September 1983.

[16] Royall, Richard, "Robust Inference Using Maximum Likelihood Estimators," Johns Hopkins University, Department of Biostatistics Working Paper 549, 1984.

[17] Smith, V.K. (ed.), *Environmental Policy Under Reagan's Executive Order* (Chapel Hill, N.C.: UNC Press, 1984).

[18] White, Halbert, "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, vol. 50 (1982) pp. 1-25.

[19] ____________ , "Corrigendum," *Econometrica*, vol. 51 (1983) p. 513.

[20] White, Lawrence, *Reforming Regulation: Processes and Problems* (Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1981).

Chapter 5

CONSTRUCTING A LIFETIME SMOKING PROFILE USING THE 1979 HEALTH INTERVIEW SURVEY

We noted above that individuals' amassed "stocks" of cigarettes consumed over a lifetime are potentially significant influences on respiratory illness. Yet the models estimated in Volume I all made use of a much more crude measure of smoking behavior. An important issue here, then, is the construction of a more sophisticated measure given available data. One theoretically plausible construct is $K(T) = \int_{\Omega} \exp(-r(T-t))C(t)dt$, where $\Omega=[\underline{T},T]$, $\underline{T}$ is time started smoking, T is present time, $C(t)$ is instantaneous cigarette consumption at t, and r is a decay or depreciation rate. The empirical representation of $K(T)$ is not straightforward, however, even given the information available in the smoking supplement to the 1979 HIS.

This is so for several reasons. First, an individual's entire lifetime smoking profile $\{C(t)\}_{t=\underline{T}}^{T}$ is never given in the data. This is so even if $C(t)$ is couched in discrete time as $\{C_t\}$ with reasonably high-frequency (e.g. one month or even one year) realizations. At best the profile can be approximated by the use of subsidiary information. Second, the above formulation is quite simple, one of an infinite number of reasonable proxies for the "true" relationship. Third, while K(T) as defined above is in principle capable of describing the effects of cigarette tar-nicotine content and cigarette length, it seems that amending the formulation to account for such influences would add little to the analysis given the nature of the data.

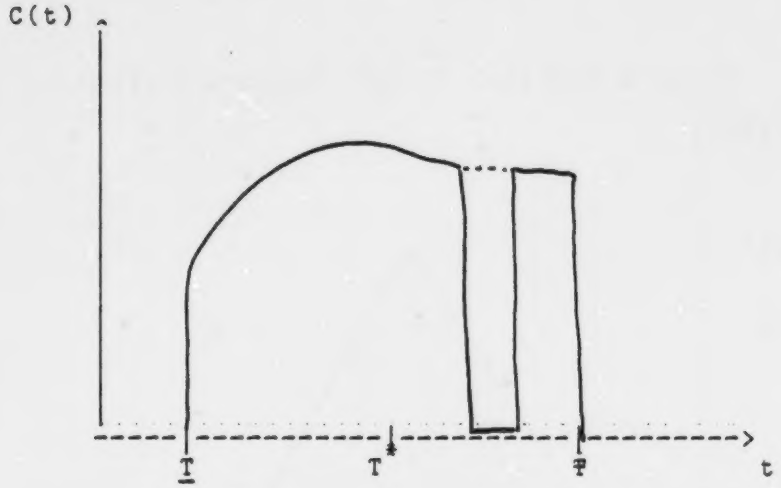
The dataset used to construct the measures is of course the 1979 HIS smoking supplement. This survey gives a reasonably detailed picture of individuals' smoking status at the time of the survey in addition to information on past attempts to quit, age at which regular cigarette smoking began, number of cigarettes smoked per day at the time of peak consumption, and other attributes. Yet most of the data in the smoking survey is of little use insofar as construction of a "packyear" or stock measure is concerned. (A check on several other datasets containing information on smoking behavior reveals similar or even more severe weaknesses.) Ignoring minor points and the complications presented by problems such as faulty recall, the most serious problems are the following.

Although data are given on peak daily cigarette consumption, no information is available on when the peak occurred (unless it coincides with present consumption levels, C(T)) nor on the duration of consumption at that peak rate. Second, information on quits (number of attempts; duration of time off) is insufficient to construct for either current or former smokers a reasonable profile of the time intervals over which C(t) was zero. Quit duration information is available only as the interval from time last smoked to T for former smokers and for the length of the single most recent quit (if any) for current smokers. Some detail is provided for current smokers on numbers of serious quit attempts, but what constitutes a "serious attempt" is analytically problematical, a subjective assessment surely varying across individuals. No information on age started smoking is given for the subsample of occasional smokers. Finally, it should be noted that even the use of an obvious stock proxy measure like $C(T-\delta)$ with, for example, δ equal one year, is precluded by data availability. It is possible to determine neither consumption levels of one year (or six months, or one month) ago nor, in many instances, even the sign of $C(T-\delta)$.

Yet there is some information that permits the construction of a reasonably interesting, albeit rough, proxy measure for the lifetime smoking profile K if one is willing to make certain assumptions. Since age started smoking is unavailable for the occasional smokers, this subsample (about two percent) will henceforth be excluded from the analysis. By assumption, $K=0$ for all never smokers. Thus, the proxy must be constructed for the subsamples of former and current smokers. The data are such that separate treatment of these two subsamples is required. In both instances, however, several plausible temporal smoking profiles can be created. In the absence of any prior information on which profile best captures an individual's true consumption path, the only sensible solution is to consider several different specifications in the empirical analysis and assess ex post the sensitivity of the results to the specification used.

For both former and current smokers, the construction of the K measures relies on a major assumption about the influence of quits on the temporal consumption profile. That is, the profile is "forgetful" of quits: once an individual resumes smoking after having quit, consumption over the quit interval is treated as if there had been no quit at all. For example, Figure 5-1 depicts the manner in which this forgetfulness operates, with true consumption $C^*(t)$ shown as a solid curve and proxy consumption shown as a dashed curve:

Figure 5-1:
Hypothetical Smoking Profile

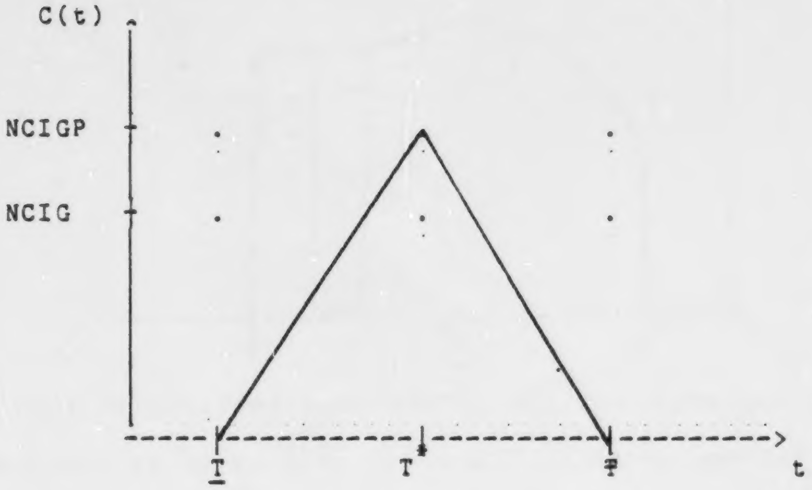


Such an approach has the unfortunate implication that, to use an extreme example, the proxy profile of an individual who quit smoking twenty years ago and resumed yesterday is drastically different than that of an individual with an otherwise identical smoking history who had not resumed smoking. Until better microdata on individuals' smoking histories become available, such drawbacks are inevitable.

For former smokers, the variables used to construct the stock proxy are time started smoking ($\underline{T}$); number of cigarettes smoked per day at peak consumption (NCIGP); and time last smoked regularly (Ŧ). There are three plausible profiles that can be constructed using this information; these can best be described graphically.

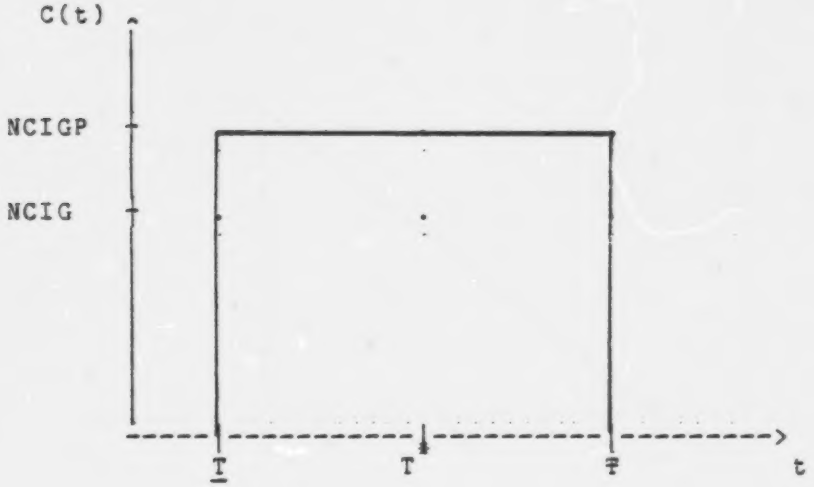
The first profile for former smokers, shown in Figure 5-2, assumes that peak consumption occurs at the midpoint ($\underline{T}$+Ŧ)/2 (=T*), and that consumption rises and falls linearly from and to zero from this peak ($C(t)$ is henceforth shown in solid lines):

Figure 5-2
Smoking Profile: Former Smokers I (F-I)



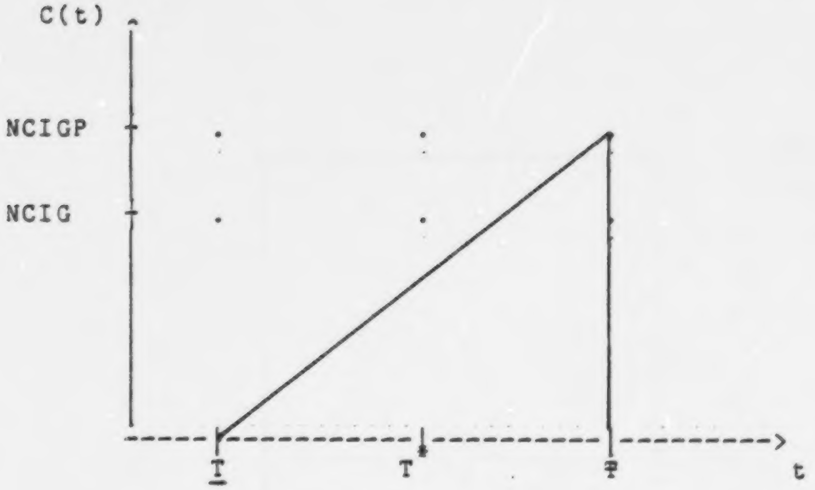
The second profile for former smokers, shown in Figure 5-3, is based on the assumption that peak consumption is attained immediately at $\underline{T}$ and continues at that rate until $\overline{T}$:

Figure 5-3
Smoking Profile: Former Smokers II (F-II)



The third former smoker profile, shown in Figure 5-4, assumes that from $\underline{T}$ consumption increases linearly to NCIGP which occurs at $\bar{T}$, then falls instantly to zero :

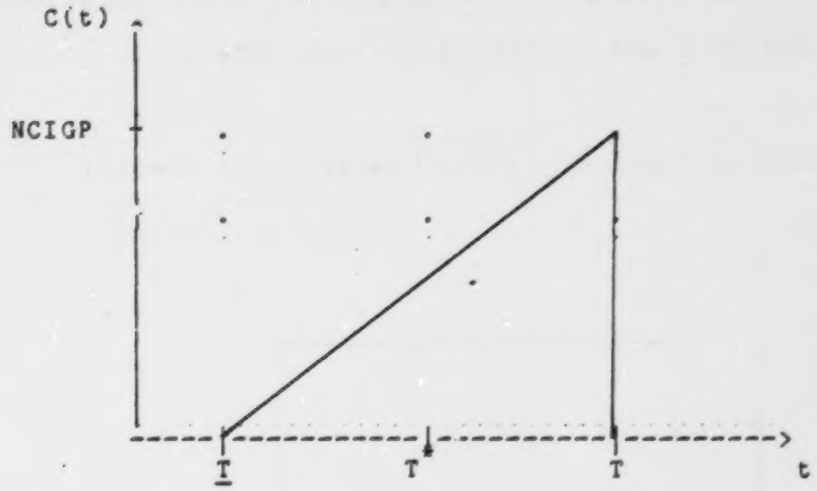
Figure 5-4
Smoking Profile: Former Smokers III (F-III)



The construction of the profiles for the current smokers uses $\underline{T}$, NCIGP, and NCIG. Five profiles seem sensible: three for current smokers for whom NCIG=NCIGP and two for those where NCIGP exceeds NCIG.

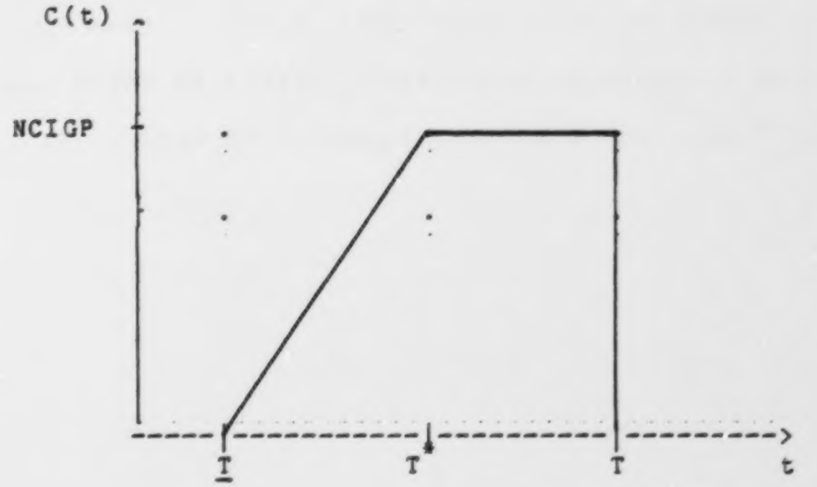
The first profile, in Figure 5-5, is analogous to that in 5-4: consumption increases linearly from $\underline{T}$ to NCIGP which coincides with NCIG at T:

Figure 5-5
Smoking Profile: Current Smokers I (C-I)



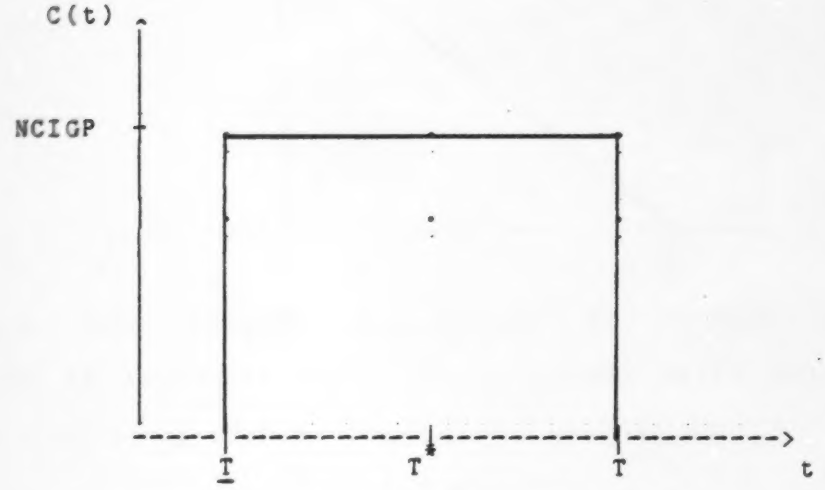
The profile in Figure 5-6 assumes that peak consumption first occurs at T*, then continues at that rate to T (T* is defined for current smokers as (T-T)/2):

Figure 5-6
Smoking Profile: Current Smokers II (C-II)



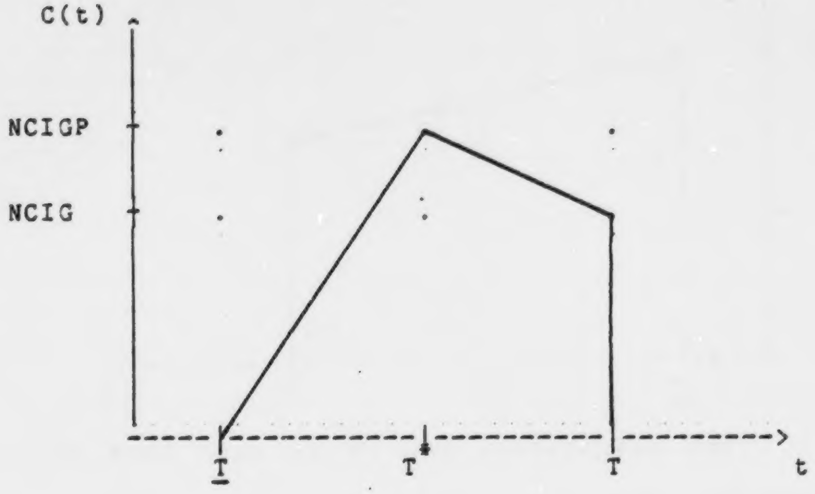
The third construct for the NCIG=NCIGP group, illustrated in Figure 5-7, assumes that NCIGP is attained immediately at $\underline{T}$ and continues at that rate to T:

Figure 5-7
Smoking Profile: Current Smokers III (C-III)



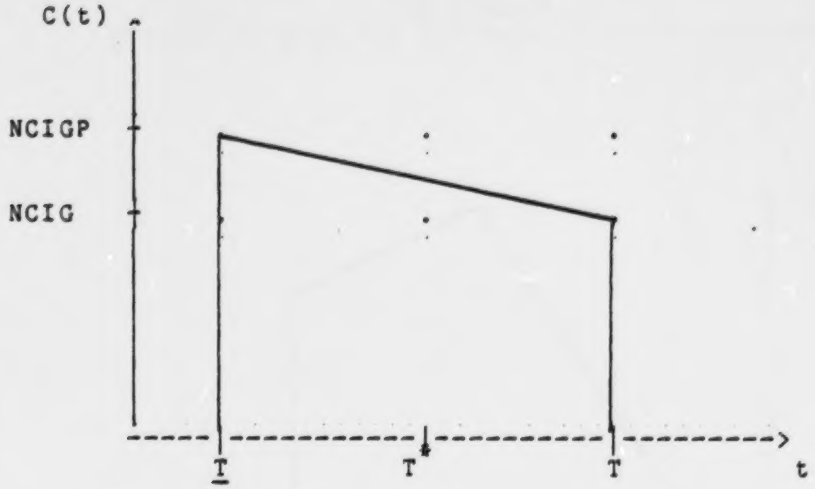
The profile in Figure 5-8 is the first shown for the subsample reporting NCIG less than NCIGP. Here it is assumed that consumption increases linearly to NCIGP which occurs at T* and then decreases linearly to NCIG at T:

Figure 5-8
Smoking Profile: Current Smokers IV (C-IV)



Finally, the profile shown in Figure 5-9 assumes that NCIGP is attained immediately at T and declines linearly to NCIG at T:

Figure 5-9
Smoking Profile: Current Smokers V (C-V)



Given these specifications, it is seen that all of the integrals to be evalated have linear or piecewise-linear integrands. That is, on the interval [a,c] (where a=$\underline{T}$ and c=$\bar{T}$ or T), the integrands are either of the form $\delta(t)(\alpha+\beta t)$, $t\epsilon[a,c]$, or $\delta(t)(\alpha_1+\beta_1 t)$, $t\epsilon[a,b]$, $\delta(t)(\alpha_2+\beta_2 t)$, $t\epsilon(b,c]$ for $b=T^*$. Specifically,

$$K(T) = \sum_{j=1}^{2} \int_{\Omega_j} \exp(-r(T-t))(\alpha_j+\beta_j t)dt,$$

where $\Omega_1=[a,b]$ and $\Omega_2=(b,c]$. Straightforward integration by parts gives the solution as

$$K(T) = \sum_{j=1}^{2} \exp(-rT)[r^{-2}(\alpha_j r+\beta_j rt-\beta_j)\exp(rt)]_{\inf(\Omega_j)}^{\sup(\Omega_j)}.$$

α_j and β_j are derived from the assumed profiles, using each individual's values for NCIG, NCIGP, $\underline{T}$, T, and $\bar{T}$.

The final point is the determination of r. Use of decay or discount rates is often essential in applied econometrics. Yet, in most instances there is no way to know the "correct" rate, so that in discounting future streams or depreciating accumulated stocks, the strategy typically adopted is to posit some rate or set of rates and conduct analysis as if the rate is known. This approach has been used in a wide spectrum of applications, generally with little discussion or justification for the rate chosen (although some studies helpfully demonstrate the sensitivity of results to the assumed rates). Such an approach will be used here.

Given the above assumptions on consumption profiles and decay rates, the K proxy measures can be derived using the relevant data in the estimation sample. However, an obvious drawback is that with three possible consumption profiles for NCIG=NCIGP current smokers, two for NCIG<NCIGP current smokers, three for former smokers, and two decay rates, there are thirty-six possible constructions of the K variable. This is an unwieldy number insofar as estimation is concerned. In order to streamline the estimation, only three combinations of the consumption profiles will be considered so that, given two decay rates, six K proxy series will be used initially for estimation. After the relative performance of these six is assessed, trimming to an even smaller number might prove expedient.

The combinations to be used are (for former, NCIG=NCIGP current, and NCIG<NCIGP smokers, respectively) {F-I, C-II, C-IV}, {F-II, C-III, C-V}, and {F-III, C-I, C-IV}. These combinations are chosen simply because of the similarity of the profiles within each combination. The K measures constructed with these combinations will be denoted K1-xx, K2-xx, and K3-xx, respectively, where xx is the decay rate (e.g. K1-05 is the K measure created using the consumption profile combination {F-I, C-II, C-IV} and the decay rate .05). Recall that C(t) is zero for all t for all never smokers, yielding K=0 for that sample.

Consider several examples of the K measure. First, a former smoker, 22 years old, who began smoking at 12 and quit smoking at 20, and who smoked 40 cigarettes per day at peak consumption: K=160. Second, a current smoker, aged 29 years, who began smoking at 19, and whose peak and current consumption is 20 cigarettes per day: K=100. Third, a former smoker, aged 83, who began smoking at 39, quit at 58, and smoked four cigarettes per day at peak consumption: K=38. Finally, a current smoker, aged 54, who began smoking at age 9, smoked three packs per day at peak, and whose current rate declined to fifteen cigarettes per day: K=1,518.75.

In the following chapter, we incorporate the smoking history measure into several models of the determinants of acute respiratory illness. The basic hypothesis here is that distinct from the risk of respiratory illness posed by exposure to ambient air pollution, *current* cigarette consumption, and other covariates, an individual's *previous* cigarette smoking predisposes him or her to an increased risk of respiratory illness. Using a subset of the proxy profiles described above enables us to test for the presence and extent of such effects.

Chapter 6

CIGARETTE SMOKING, AIR POLLUTION, AND RESPIRATORY ILLNESS: AN ANALYSIS OF RELATIVE RISKS

6.1 Introduction

The relative risks associated with cigarette smoking and ambient air pollution are difficult to assess. First, individuals' health status is largely subjective and often difficult to measure.[1] In addition, lung physiology is complex as well as heterogeneous in a population of individuals, hindering both the identification and measurement of all potential determinants of respiratory illness. Moreover, there is little theoretical guidance as to the likely form of any functional relationship between risk exposure and illness response. Finally, data on exposure to risks are often not all one might like them to be. It is thus apparent why one expert on quantitative risk assessment was moved to comment:

> Quantitative risk assessment is not a panacea. A primary limitation is that such an assessment is concerned only with what can be measured and quantified.[2]

In spite of these problems, relative risk assessments must be undertaken for smoking and air pollution. Both have been the subject of much discussion and study in the health and environmental policy communities, and in the popular press as well. Moreover, as discussed below, considerable resources are being devoted to understanding and reducing both risks. It is essential, therefore, that the risks attributable to smoking and air pollution be assessed simultaneously within a single coherent framework; otherwise, risks attributed to one may in fact be due to the other, thus biasing any estimates of the health risk of but one of the variables.

The plan in this chapter is as follows. Section 2 discusses in greater detail the problem of acute respiratory illness and some possible links between it and smoking and air pollution. Section 3 describes the dataset used in the empirical analysis, explains the health measures utilized, and sketches the estimation strategy. Finally, Section 4 presents empirical results, derives the estimates of the relative risks of interest, and briefly suggests new directions for future research in this area.

6.2 Smoking, Pollution, and Acute Illness

The association between cigarette smoking and several major chronic illnesses is well known. The 1982, 1983, and 1984 reports of the U.S. Surgeon General detail and publicize, respectively, the relationships between cigarette smoking and cancer, cardiovascular disease, and chronic obstructive lung disease. For these diseases, the indictment of cigarette smoking is strong: although data can obviously never demonstrate causality (as the Tobacco Institute is wont to remind), the correlative evidence is overwhelming.

Less widely publicized are associations between cigarette smoking and less severe illnesses. There is evidence, however, that suggests the existence of such linkages. Chapters Three and Six of the 1979 Surgeon General's report summarize much of the existing research in this area. There it is reported that relative to nonsmokers, current smokers have more frequent respiratory tract infections and a greater prevalence of cough; symptoms like cough and sputum production tend to increase with the number of cigarettes smoked. Moreover, the 1979 report finds that "...people who had ever smoked...had a higher incidence of acute illnesses than did people who had never smoked (p. 3-6)." Smokers report approximately 45% more illness-related work loss days than do never smokers.

Owing to the magnitude and severity of the illnesses associated with cigarette smoking, considerable attention and public resources have been devoted to the study of and remedies for such illnesses. While one is hard-pressed to estimate the value of the resources spent in such activities, it is safe to venture that the value is enormous.

Other public policies have been put in place to protect individuals' respiratory health. For example, several federal agencies are involved in the protection against and compensation for damages from pneumoconiosis (black lung disease), while exposure to respirable hazardous substances in the workplace -- like the cotton dust which causes byssinosis -- comes under the regulatory purview of OSHA.

Of more immediate interest here, however, is the widespread concern that ambient air pollution may be detrimental to individuals' respiratory health. Many clinical and epidemiological studies have tested for possible relationships between ambient air pollution and both morbidity and mortality. The cornerstone of federal air pollution policy in the U.S., the Clean Air Act, places primary emphasis on the protection of public health from air pollution, insisting that air quality standards be set to provide "an adequate margin of safety...to protect the public health." Of the possible air pollution-related illnesses, it is of course respiratory illness that is of utmost concern. Regulatory mandates pursuant to the Clean Air Act are not inexpensive: according to the most recent estimates, annual costs of complying with the Act are approximately $25 billion. This sum is sure to grow as older sources of pollution are retired and newer ones -- which must meet stricter emissions standards -- are built to replace them. Whether such expenditures achieve desired ends efficiently (if at all) is a question on which we hope to shed some light.

Our task in this chapter is to assess the relative contributions of cigarette smoking and air pollution to an individual's risk of suffering from acute respiratory impairments of varying severity. We concentrate on this category because none but the most extraordinary exposures to air pollution can be expected to rival direct (or perhaps even passive) smoking as a cause of the more serious illnesses like cancer, cardiovascular disease, or chronic lung disease. It seems to us plausible to hypothesize that if typical levels of ambient air pollution in the U.S. are to influence individuals' health in any manner, then acute respiratory illness must be a primary area of suspicion. Because it can be quite expensive to control air pollution, it is important to assess how the benefits of doing so compare to those associated with policies oriented towards smoking cessation. We suggest that an analysis of the relative contributions of cigarette smoking and ambient air pollution to acute respiratory illness is one way to approach this important assessment.[3]

6.3 Data and Estimation Strategy

The individual data used here are from the 1979

Health Interview Survey (HIS), a national sample of approximately 110,000 individuals conducted over the course of each year by the National Center for Health Statistics. In this regard, the analysis in this chapter is similar to that in Volume I and in Chapter 4 in the present volume. The socioeconomic data elicited from each respondent in the HIS includes information on age, race, sex, income, education, as well as other individual- and household-specific characteristics. In addition, the supplemental survey on smoking behavior administered in the 1979 HIS make it particularly useful for present purposes. This supplemental questionnaire was asked of one-third of the approximately 78,000 adults (17+ years) interviewed, and provides detailed data on lifetime smoking history and present smoking behavior.

Restrictions in activity due to any illness experienced during the two-week period prior to the date of each interview are reported by the interviewee or another household member responding for the interviewee. Manifestations of illnesses are classified in three types: bed disability days, work or school loss days, and what might best be thought of as minor restricted activity days. The latter are days on which the subject was neither bedridden nor forced to miss work or school, but on which the individual did suffer from an impairment sufficient to cause a perceptible restriction on usual activity. The information on health impairments elicited in the survey is coded by cause according to the International Classification of Disease. As discussed earlier, attention is limited in this chapter to those restrictions in activity due to respiratory illness.

All air pollution data come from the U.S. Environmental Protection Agency's SAROAD system. The air quality data used here are measured over the two-week recall period for which the individual acute health data are available. The received opinion of respiratory physiologists suggests that not all airborne pollutants are equally important in influencing respiratory health. Accordingly, the analysis of acute respiratory illness here uses air pollution data for ozone (OZONE), a gaseous pollutant that is the primary constituent of smog, and sulfates (SULFATE), perhaps the most harmful of airborne particulate matter. The subsequent analysis characterizes individuals' exposures to air pollution using data from the air pollution monitors nearest the center of the census tract in which the individual resides. No individual for whom the nearest monitor is more than ten miles away is included in the estimation sample, with the sample average distance from centroid to monitor being slightly more than four miles. For more details on the air pollution data used in this analysis, consult Volume I.

We include two measures to control for cigarette smoking. The first is the individual's daily consumption of cigarettes at the time of the interview (NCIG). (The HIS unfortunately contains no information on cigar or pipe smoking.) Since the consumption data are self-reported, some caution must be exercised in light of Warner's underreporting hypothesis [5]; however, we make no attempt here to correct for this possible errors-in-variables problem. (Interestingly, in light of the mounting evidence on the harms of passive smoking, attributing zero as the number of cigarettes smoked per day by a "nonsmoker" represents perhaps an understatement of the daily dosage of cigarettes.)

Both medical evidence and common sense suggest that the rate of current cigarette consumption alone is an insufficient characterization of an individual's smoking-related risk of respiratory illness (see, for example, Chapter Six of the 1979 Surgeon General's report). A more appropriate characterization of these risks incorporates the influences of both current as well as past cigarette consumption. Accordingly, the influence on the likelihood of current acute respiratory illness of lifetime cigarette consumption is measured by the variable PACKS, a proxy for the number of cigarette packs that a given individual has "amassed" over his or her lifetime. PACKS can be viewed as a stock or state variable equal to the integral over an individual's lifetime cigarette pack consumption profile {C(t)}. (See Chapter 5 for a discussion of the creation of the K(t) measures.) The measure defined in Chapter 5 and converted into pack units is selected from the set of candidates to serve as the packyear proxy in the present analysis.

Table 6-1 provides a summary description of the air pollution and smoking measures, as well as the other independent variables used, and Table 6-2 depicts their sample means, minima, and maxima.

Among the measures of respiratory illness available in the HIS, the number restricted activity days due to respiratory illness during the two-week recall period (RRAD) is a logical choice for use in the present analysis. However, one drawback to its use as a measure of health status is that it is a somewhat aggregated concept. Any day reported as a bed disability or work loss day, when due to respiratory illness, is counted as a RRAD, as are days when individuals are hampered in minor ways from performing usual activities without confinement to bed or work loss. It is possible, however, that the determinants of minor restrictions are likely to be different -- in kind or in magnitude -- from the determinants of severe limitations.

The HIS data do not enable a complete disaggregation of these different types of respiratory restrictions. For

Table 6-1

Variable Definitions

Variable Name	Description
OZONE	Average daily maximum one-hour ozone reading during two-week recall period at monitor nearest the centroid of respondent's census tract of residence, subject to ten-mile distance cutoff (in parts per million)
SULFATE	Average 24-hour sulfate concentration during two-week period at monitor nearest the centroid of respondent's census tract of residence, subject to ten-mile distance cutoff (in $\mu g/m^3$)
NCIG	Number of cigarettes smoked per day
PACKS	Proxy for lifetime cigarette consumption, in packs (see text or [5] for detailed description)
TEMP	Average daily maximum temperature during two-week recall period (in degrees F)
RAIN	Average daily precipitation during two-week recall period (in inches)
AGE	Age, in years
EDUC	Number of years of schooling
INCOME	Annual family income, in 1979 dollars
CHRONLIM	Equals 1 if respondent reports a persistent limitation in activity due to a chronic ailment, equals 0 otherwise
MALE	Equals 1 if male, equals 0 if female
WHITE	Equals 1 if white, equals 0 if black
BLUECOL	Equals 1 if respondent reports usual activity is working and usual employment is blue collar, equals 0 otherwise
WHITECOL	Equals 1 if respondent reports usual activity is working and usual employment is white collar, equals 0 otherwise
INSCHOOL	Equals 1 if respondent reports usual activity is going to school, equals 0 otherwise

Table 6-2

Sample Summary of Independent Variables (n=3073)

Variable	Mean	Minimum	Maximum
OZONE	.0426	0	.2136
SULFATE	10.87	.784	52.14
NCIG	7.454	0	98
PACKS	3299.1	0	44174.1
TEMP	63.92	11.14	106.36
RAIN	.114	0	.637
AGE	42.83	17	96
EDUC	11.73	0	18
INCOME	17095	500	30000
CHRONLIM	.173	0	1
MALE	.433	0	1
WHITE	.857	0	1
BLUECOL	.232	0	1
WHITECOL	.288	0	1
INSCHOOL	.079	0	1

example, due to some peculiarities in data collection, it is not possible in all instances to disentangle work loss from bed restricted days (see Volume I for a detailed discussion of these problems). However, the data do permit a unique disaggregation of RRADs into two qualitatively distinct types: days restricted in activity due to respiratory illness but not confined to bed (minor RRAD, or RADM), and days restricted in activity due to respiratory illness with bed confinement (severe RRAD, or RADS). Thus, it is possible to determine for each individual the number of RADM, RADS, and nonrestricted days (NRD) occurring during the two-week recall period. In the analysis to follow, we consider two separate definitions of RADM and RADS: first, those related to respiratory illness classified by NCHS as either chronic or acute (RADM-CA, RADS-CA), and, second, those related only to respiratory illness classified as acute (RADM-A, RADS-A).* The sample frequencies are presented in Table 6-3.

The nature of these health status measures is such that several peculiar characteristics must be treated simultaneously in the estimation procedure. First, the measure best suited for the analysis is multivariate in nature: during any two-week period, individuals can report minor or severe restrictions in activity due to respiratory illness, or can report no respiratory

Table 6-3

Sample Frequency Distribution of RRAD Measures (n=3073)

Number of Days	RADM-CA	RADM-A	RADS-CA	RADS-A
0	3013	3027	3007	3014
1	16	14	19	18
2	14	12	18	15
3	8	6	10	10
4	6	5	5	5
5	2	2	4	4
6	1	1	1	1
7	2	1	3	2
8	1	1	3	3
9	1	1	0	0
10	1	1	0	0
11	0	0	0	0
12	1	1	0	0
13	0	0	0	0
14	7	1	3	1

impairment. Second, outcomes are mutually exclusive. On a day where an individual reports a RADM, neither a RADS nor a NRD can be reported; similar exclusivity holds for RADS and NRD. Third, for all individuals, each of RADM, RADS, and NRD is constrained to take integer values in {0,1,...,14}, with the sum RADM+RADS+NRD equal to fourteen. Finally, because of the protocol of the HIS, it is not possible to determine on what days during the two-week recall period a given individual reported the RADM, RADS, or NRD; only the number of each type of outcome is known. While it seems sensible to suppose that RRADs would be contiguous rather than disparate during any particular time interval, the data used here do not permit such a conjecture to be verified.

Following the discussion in Chapter 2 of this volume, the estimation strategy is to view each day during the two-week recall period as a trial on which one and only one of the three possible outcomes can occur. For each individual, then, there are fourteen trials. Because any one individual's covariates are invariant across the fourteen trials and, as noted above, because it is impossible to ascertain which health outcomes occurred on which days (except, of course, in the polar case where the same outcome occurs on all fourteen days), it is plausible for estimation purposes to assume independence both across trials for an individual and across individuals. (In the estimation subsample used, it happens that at most one individual per household is included. Thus, contagion effects -- which might otherwise vitiate the assumption of independence across individuals -- can be ignored.)

The preceeding paragraphs describe a model that can be appropriately cast in terms of a multinomial distribution with $k=3$ possible outcomes; $n_t=n_\tau=n=14$ independent trials for all t,τ; and probability vector $(\pi_{M_t},\pi_{S_t},\pi_{N_t})$ (M=RADM, S=RADS, N=NRD) such that $\pi_{M_t}+\pi_{S_t}+\pi_{N_t}=1$. The number of successes or incidences of each type is n_{q_t} for $q=M,S,N$, and $n_{M_t}+n_{S_t}+n_{N_t}=14$ for all t. Thus, denoting the multinomial (vector) random variable for the t-th individual as Y_t,

$$Pr(Y_t=y_t) = n! \prod_{q\in\Omega} [(\pi_{q_t})^{n_{q_t}}]/n_{q_t}!, \qquad (1)$$

where $\Omega=\{M,S,N\}$ and $n=14$. A logit specification for the π_{q_t} is assumed:

$$\pi_{q_t} = \exp(X_t\beta_q)/(\sum_{r\in\Omega}\exp(X_t\beta_r)), \qquad (2)$$

for $q=M,S,N$. The parameter vectors β_q are unique only up to a difference, so that some normalization is necessary; $\beta_N=0$ is used here. Details on estimation are presented in the appendix.

A basic and more popular version of the model described above is the ordered logit model described in Chapter 2, in which it is assumed that there exists some mechanism that orders the outcome probabilities according to a particular latent measure (illness severity, for example). The typical assumption is that the coefficients $\beta^*_q = (\beta_{q2}, \dots, \beta_{qK})$ are invariant across the outcomes q (except for the outcome whose parameter vector remains normalized to zero), with the ordering is characterized by outcome-specific intercept terms, such that $\beta_{p1} < \beta_{q1} < \beta_{r1} < \beta_{s1} \dots$. It is not unreasonable to hypothesize that the respiratory-related restricted activity day health outcome probabilities might follow such an ordering, with RADS → RADM → NRD, "→" signifying the ordering "more severe than." For purposes of comparison, therefore, we also present estimates of a a multiple-trial version of an ordered logit model. It happens that this is a parameter-restricted version of the multinomial model specified above, with (K-1) restrictions of the form $\beta^*_M = \beta^*_S$ on the likelihood function (A.2) implied by the ordered logit likelihood function. It is thus possible to test in a straightforward manner whether these restrictions are valid insofar as the model and data sample used here are concerned. The test is a standard likelihood ratio test, with the test statistic computed as $LR = -2(\ell_O - \ell_M)$; ℓ_O is the maximized likelihood function value for the ordered logit specification and ℓ_M is the corresponding value for the multinomial model (A.2). Under the null hypothesis that the (K-1) restrictions are valid, LR is distributed asymptotically as central χ^2 with (K-1) degrees of freedom.

6.4 Estimates of Model Parameters and Relative Risks

The estimates of the model using the chronic and acute RRAD measures of respiratory illness are presented in Table 6-4. Insofar as the parameter estimates associated with the independent variables other than smoking or pollution are concerned, it is seen that most are statistically significant in at least one of the RADM-CA or RADS-CA estimated parameter vectors, with generally plausible signs in most instances. The parameter estimate associated with the current level of cigarette smoking (NCIG) is statistically important in β_M but is insignificant in β_S. Lifetime cigarette consumption (PACKS) plays an opposite role: its associated parameter estimates are positive and significant in the β_M vector, but statistically indistinguishable from zero in β_S. SULFATE appears to be an insignificant contributor to either RADM-CA or RADS-CA. OZONE, conversely, has an associated parameter estimate in β_M that is positive and statistically significant, although the ozone coefficient in β_S is statistically unimportant. This finding is consistent with those in Volume I. There, using more

Table 6-4
Model Estimates: Chronic and Acute RRADs with Linear Risk Factor Influence

Variable	β_{M-CA}	β_{S-CA}
INTERCEPT	-6.34	-3.64
	(9.8)	(7.2)
OZONE	7.03	1.77
	(2.5)	(.45)
SULFATE	.0061	-.012
	(.54)	(.84)
NCIG	-.0034	.026
	(.61)	(4.8)
PACKS	.36E-4	.53E-5
	(3.1)	(.33)
TEMP	-.014	-.025
	(3.4)	(5.2)
RAIN	0.85	0.45
	(1.3)	(.62)
AGE	-.0084	.31E-3
	(1.9)	(.061)
EDUC	-.0082	-.058
	(.39)	(2.4)
INCOME	-.32E-4	-.63E-4
	(4.1)	(7.0)
CHRONLIM	0.99	0.46
	(6.7)	(2.6)
MALE	0.28	.079
	(2.1)	(.53)
WHITE	2.59	0.78
	(5.1)	(3.2)
BLUECOL	-1.76	-0.17
	(5.9)	(.75)
WHITECOL	-0.18	0.83
	(1.0)	(4.3)
INSCHOOL	-0.42	0.90
	(1.4)	(3.1)

$\text{Log}(\ell) = -2734.55$

Note: Asymptotic normal scores for $H_0: \beta_k = 0$ in parentheses.

"primitive" OLS and logit techniques, we found positive and often significant associations between ozone and minor illnesses among adults, but no pattern of associations when we examined either work loss or bed disability days. Thus, the findings in this chapter provide some corroborative evidence using a more sophisticated and appropriate statistical approach.

Similar estimates obtain in the model of the acute-only respiratory ailments RADM-A and RADS-A, presented in Table 6-5. Most notable is that the individual parameter significance levels tend to be somewhat lower than those estimated in the chronic-acute model of Table 6-4, although the qualitative interpretation is in most instances unchanged. Of particular import is that the coefficient estimates associated with OZONE and PACKS in β_M are no longer significant at the 95% level.[3]

In Chapter 4 we found that various nonlinear transformations of the ozone measure lead to differing conclusions about the significance of the relationship between ozone and respiratory health. There, remember, the transformation $(\text{OZONE})^{.5}$ proved most significant. As such nonlinearities are potentially important in the present analysis as well, we also consider simple transformations of OZONE, NCIG, and PACKS of the form OZONE^{λ_1}, NCIG^{λ_2}, and PACKS^{λ_3} for $\lambda_1, \lambda_2, \lambda_3 > 0$. (We ignore

Table 6-5
Model Estimates: Acute-only RRADs with Linear Risk Factor Influence

Variable	β_{M-A}	β_{S-A}
INTERCEPT	-6.67	-4.66
	(8.9)	(8.1)
OZONE	5.09	-2.39
	(1.3)	(.49)
SULFATE	.0081	.0031
	(.59)	(.22)
NCIG	-.0088	.023
	(1.1)	(3.5)
PACKS	3.1E-4	1.4E-5
	(1.8)	(.072)
TEMP	-.0069	-.024
	(1.2)	(4.4)
RAIN	0.46	0.54
	(.53)	(.68)
AGE	.0053	.0057
	(.95)	(.98)
EDUC	-.0088	-.039
	(.31)	(1.4)
INCOME	.28E-4	.41E-4
	(2.8)	(4.4)
CHRONLIM	-1.10	0.13
	(3.6)	(.58)
MALE	-0.15	-0.30
	(.79)	(1.8)
WHITE	1.82	0.65
	(3.6)	(2.4)
BLUECOL	-1.25	0.55
	(3.6)	(2.1)
WHITECOL	0.13	1.35
	(.59)	(5.9)
INSCHOOL	0.28	1.57
	(.82)	(5.0)

$\log(\ell) = -2048.53$

Note: Asymptotic normal scores for $H_0: \beta_k = 0$ in parentheses.

transformations of SULFATE because of the its generally insignificant contributions as witnessed in Tables 6-4 and 6-5.) The software used for estimation does not enable maximum likelihood estimation of the λ_j; instead, a grid search approach is used, where the search is over $(\lambda_1,\lambda_2,\lambda_3) \varepsilon \Psi \times \Psi \times \Psi$, and $\Psi = \{0.5, 1.0, 1.5, 2.0\}$. Of the sixty-four possible $(\lambda_1,\lambda_2,\lambda_3)$ triples, that which maximizes the conditional (on λ_1,λ_2) likelihood function with respect to (β_M,β_S) is selected as the (pseudo) MLE.

The estimates of the RADM-CA and RADS-CA model using the nonlinear transformations are presented in Table 6-6. The pseudo-MLEs of the λ_k are $\lambda_1 = 0.5$, $\lambda_2 = 1.5$, and $\lambda_3 = 1.0$, with a likelihood ratio test indicating that these transformations are jointly significant at greater than the 95% level. The overall qualitative findings are unchanged; however, the parameter estimates associated with the transformed risk factors are more finely resolved than those presented in Table 6-4. Similar statements can be made about the acute-only model, whose estimates are presented in Table 6-7. Again, the pseudo-MLEs for the λ_k are 0.5, 1.5, and 1.0 for the OZONE, NCIG, and PACKS transformations, respectively. The likelihood ratio test of the joint significance of the transformations is significant only at slightly above the 90% level; since the λ_k are not true MLEs, however, such an LR test is somewhat misleading, and is biased in favor of accepting

Table 6-6
Model Estimates: Chronic and Acute RRADs with Nonlinear Risk Factor Influences

Variable	β_{M-CA}	β_{S-CA}
INTERCEPT	-6.75 (10.5)	-3.64 (7.2)
OZONE$^{\lambda_1}$	5.10 (3.5)	2.86 (1.7)
SULFATE	.0039 (.35)	-.014 (1.0)
NCIG$^{\lambda_2}$	-.25E-3 (.32)	.0034 (5.3)
PACKS$^{\lambda_3}$	.35E-4 (3.0)	.81E-5 (.53)
TEMP	-.020 (4.2)	-.030 (5.9)
RAIN	1.01 (1.5)	0.57 (.78)
AGE	-.0080 (1.9)	-.94E-3 (.19)
EDUC	-.0071 (.33)	-.059 (2.4)
INCOME	-.32E-4 (4.1)	-.64E-4 (7.1)
CHRONLIM	0.99 (6.7)	0.46 (2.9)
MALE	0.28 (2.0)	.064 (.43)
WHITE	2.56 (5.0)	0.75 (3.0)
BLUECOL	-1.77 (6.0)	-0.15 (.65)
WHITECOL	-0.19 (1.1)	0.86 (4.4)
INSCHOOL	-0.41 (1.3)	0.86 (3.0)

$\mathrm{Log}(\ell)=-2729.13$

Note: Asymptotic normal scores for $H_0:\beta_k=0$ in parentheses.
λ_k from grid search: $\lambda_1=0.5$; $\lambda_2=1.5$; $\lambda_3=1.0$.

Table 6-7
Model Estimates: Acute-only RRADs with Nonlinear Risk Factor Influence

Variable	β_{M-A}	β_{S-A}
INTERCEPT	-7.09 (9.5)	-4.59 (7.9)
OZONE$^{\lambda_1}$	4.89 (2.5)	1.35 (.71)
SULFATE	.0052 (.37)	.44E-3 (.032)
NCIG$^{\lambda_2}$	-.33E-3 (.26)	.0034 (4.1)
PACKS$^{\lambda_3}$	.23E-4 (1.3)	-.88E-6 (.046)
TEMP	-.013 (2.1)	-.029 (5.2)
RAIN	0.68 (.78)	0.69 (.87)
AGE	.0066 (1.2)	.0052 (.91)
EDUC	-.0076 (.27)	-.040 (1.4)
INCOME	-.28E-4 (2.8)	-.42E-4 (4.5)
CHRONLIM	-1.09 (3.6)	0.13 (.58)
MALE	-0.14 (.74)	-0.31 (1.9)
WHITE	1.79 (3.5)	0.62 (2.3)
BLUECOL	-1.27 (3.7)	0.56 (2.2)
WHITECOL	0.12 (.54)	1.36 (6.0)
INSCHOOL	0.30 (.89)	1.55 (4.9)

Log(ℓ)=-2045.12

Note: Asymptotic normal scores for $H_0:\beta_k=0$ in parentheses.
λ_k from grid search: $\lambda_1=0.5$; $\lambda_2=1.5$; $\lambda_3=1.0$.

the null that transformations are not important.

The estimates of the ordered logit models are presented in Table 6-8. The λ_k transformations suggested by the multinomial models of both Table 6-6 and Table 6-7 are used here. In the model of RADM-CA and RADS-CA (column 1), the parameter estimates for $(OZONE)^{.5}$, TEMP, INCOME, CHRONLIM, WHITE, BLUECOL, $(NCIG)^{1.5}$ and PACKS are all significant at greater than the 99% level. A perhaps peculiar result is that the estimate of the RADM intercept exceeds that for RADS, thus calling into question the validity of the ordered logit specification. Indeed, the $\chi^2_{(15)}$-distributed likelihood ratio test statistic of the restrictions on the multinomial model that are implied by the ordered specification has a value of 89.58, suggesting that the ordered specification can be rejected with considerable confidence in favor of the unrestricted multinomial model. Similar results obtain for the model of RADM-A and RADS-A (column 2): estimated parameters associated with TEMP, INCOME, WHITE, WHITECOL, INSCHOOL, and $(NCIG)^{1.5}$ are significant at above the 99% critical level, while the asymptotic t-statistics associated with $(OZONE)^{.5}$ and CHRONLIM parameters exceed the 95% level. In this instance, the χ^2 test statistic for the ordered logit model restrictions has a value of 68.58, again suggesting that the ordered specification be rejected in favor of the general multinomial model.

Table 6-8
Model Estimates: Ordered Logit, Nonlinear Risk Factor Influence

Variable	Chronic-Acute	Acute-only
INTERCEPT-RADM	-4.28 (11.6)	-4.82 (11.0)
INTERCEPT-RADS	-5.07 (13.7)	-5.45 (12.3)
OZONE$^{\lambda_1}$	4.17 (3.8)	3.15 (2.3)
SULFATE	-.0033 (.37)	.0033 (.35)
NCIG$^{\lambda_2}$	.0017 (3.4)	.0020 (3.0)
PACKS$^{\lambda_3}$	.24E-4 (2.6)	.93E-4 (.71)
TEMP	-.024 (7.1)	-.022 (5.3)
RAIN	0.76 (1.5)	0.57 (1.1)
AGE	-.0045 (1.4)	.0063 (1.6)
EDUC	-.029 (1.8)	-.024 (1.2)
INCOME	-.46E-4 (7.9)	-.35E-4 (5.2)
CHRONLIM	0.75 (6.7)	-0.41 (2.3)
MALE	0.17 (1.7)	-0.22 (1.8)
WHITE	1.38 (6.3)	0.99 (4.2)
BLUECOL	-0.93 (5.3)	-0.26 (1.3)
WHITECOL	0.28 (2.1)	0.71 (4.6)
INSCHOOL	0.21 (1.0)	0.92 (4.0)
Log(ℓ) =	-2773.84	-2079.41

Note: Asymptotic normal scores for $H_0: \beta_k = 0$ in parentheses.
λ_k from grid search: $\lambda_1 = 0.5$; $\lambda_2 = 1.5$; $\lambda_3 = 1.0$.

On the basis of these results, we elect to use the estimates of the nonlinear risk factor multinomial models presented in Tables 6-6 and 6-7 as the foundation of the relative risk estimates. In the multinomial model, translations from the qualitative outcomes to quantitative estimates of relative risks are fairly straightforward. One obvious strategy for evaluating the relative risks of smoking and air pollution would be to assess and compare the estimated elasticities of the daily outcome-specific probabilities with respect to the pollution and smoking control variables. Using the incidence probabilities defined in (2), and allowing for the cases where the control variables are subject to the nonlinear transformations $h(x_{tk})=(x_{tk})^{\lambda_k}$, the elasticity formula is:

$$\varepsilon_{\pi_{qt},x_{tk}} =$$

$$\lambda_k (x_{tk})^{\lambda_k} [\beta_{qk} - (\sum_{r\varepsilon\Omega} \exp(X_t^h \beta_r)\beta_{rk}) / (\sum_{r\varepsilon\Omega} \exp(X_t^h \beta_r)), \quad (3)$$

which simplifies when $\lambda_k=1$. (Here, X_t^h denotes the transformed X_t vector.)

While the elasticity comparison approach provides perhaps the most straightforward method for assessing the relative risks of interest, the nature of the data used here renders it somewhat uninformative. In brief, the problem is that 64% of the sample are classified as current nonsmokers, while 44% are never smokers. It is seen by inspection of equation (3) that for these large subsamples the estimated NCIG and PACKS elasticities are zero.

We therefore adopt in lieu of elasticity comparison an approach that considers the discrete changes from the baseline or prevailing daily incidence probabilites attributable to a variety of discrete changes in the control variables from their prevailing sample values. This strategy has at least two advantages. First, it circumvents the non- or never-smoker problem. Second, the magnitudes of the hypothetical discrete changes in the control variables are set to mimic potentially interesting policy measures.

The strategy is as follows. First, for each of the four incidence outcomes (RADM-CA, RADM-A, RADS-CA, RADS-A), a baseline mean probability is calculated using the estimated models in Tables 6-7 and 6-8. This mean probability -- which is simply the sample average of the $\hat{\pi}_{q_t}$ -- is denoted π_q^{α}. The second step is to perturb the control variable of interest in each X_t by the specified amount and reevaluate each individual's incidence probability using the perturbed X_t. The sample average of these new $\hat{\pi}_{q_t}$ is denoted π_q^{ω}. Finally, for each illness measure and each control perturbation considered, the diffference $(\pi_q^{\omega}-\pi_q^{\alpha})$ is calculated. The results are presented in Table 6-9.

As Table 6-9 indicates, depending upon the specific model of interest, changes in either or both measures of smoking as well as ambient ozone concentrations can affect the likelihood of an individual's reporting a minor or severe respiratory impairment. For instance, in the RADM-CA model a 5 percent increase from the sample mean in the average daily maximum 1-hour ozone concentration increases the estimated risk of a minor respiratory impairment on any given day by an average of $1.76*10^{-4}$. A comparable increase in risk is predicted to result from an individual's having smoked slightly more than an incremental one pack per day for two years (since one pack per day for one year adds $0.75*10^{-4}$ to the risk of a RADM-CA). The same model reveals that a ten percent increase in the average daily maximum ozone concentration is about equivalent to an increase of an extra one pack a day smoked for five years in terms of incremental risk ($3.50*10^{-4}$ and $3.85*10^{-4}$, respectively).

We can also compare the incremental risks of air pollution with those associated with current cigarette consumption. For instance, from the RADS-CA model, a five percent increase in ambient ozone concentrations increases the baseline risk of a severe acute respiratory illness by $0.89*10^{-4}$. This is about one-twelfth the effect of an individual's currently smoking an additional half pack of

Table 6-9

Estimated Mean Changes from Baseline Probabilities π_q (x10,000)

	RADM-CA	RADS-CA	RADM-A	RADS-A
Baseline π_q	60.202	50.671	35.331	40.910
Hypothetical Control Change:				
$OZONE_t + .05 * \overline{OZONE}$	1.76	0.89	0.98	0.35
$OZONE_t + .10 * \overline{OZONE}$	3.50	1.75	1.95	0.68
$NCIG_t + 5$	--	4.47	--	3.46
$NCIG_t + 10$	--	10.80	--	8.40
$PACKS_t + 365$	0.75	0.14	0.29	--
$PACKS_t + 1825$	3.85	0.71	1.49	--

Notes: "--" signifies negative predicted change

$\overline{OZONE}$ signifies the sample mean concentration of $OZONE_t$, = .0426 cigarettes per day. Comparable calculations could be made in the other models as well (although we prefer not to go into detail here since the significance levels on the variables of interest are not sufficiently high to warrant large confidence in the estimated risk changes).

For purposes of public policy, it would be desirable to go beyond the estimation of relative risks to consider the cost and efficacy of "control" measures. This would permit at least crude cost-effectiveness comparisons to be made. Some estimates are available regarding ozone reductions. White [6] reports that when the National Ambient Air Quality Standard for ozone was reviewed in 1978, the marginal cost of meeting a standard of 0.12 ppm as opposed to one of 0.14 ppm was approximately $2.0 billion. Although the form of that standard (second highest hourly reading at a monitor) differs from the measurement of ozone in this study (average daily maximum one-hour reading during a two-week period), a link between the two could be made. This would permit an estimate of the costs per unit of predicted ozone risk reduction, holding other possibly beneficial effects of ozone reduction (agricultural productivity increases, for example) constant.

In principle, estimates could be assembled on the costs of reducing cigarette consumption (for an excellent discussion of the nature of such costs, see [1]). Using such data, and the results presented above, estimates of cost per unit of reduced risk from smoking could be derived and compared with those resulting from pollution control. Finally, if appropriate allowances were made for the qualitatively different nature of the two risks -- the differing degrees of voluntarism, for instance -- it would be possible to draw inferences about potentially efficient resource allocation.

APPENDIX

Given an independent sample of T observations, the likelihood function is

$$L = n! \prod_{t=1}^{T} \{ \prod_{r \in \Omega} [(\pi_{r_t})^{n_{r_t}}]/n_{r_t}! \}. \qquad (A.1)$$

In logs,

$$\ell = \sum_{t=1}^{T} \sum_{r \in \Omega} \{ n_{r_t} [X_t \beta_r - \log(\sum_{s \in \Omega} \exp(X_t \beta_s))] \} + c, \qquad (A.2)$$

where c does not depend on $\beta=(\beta_M,\beta_S)$. ℓ is concave in β, thus assuring convergence.

A Newton-Raphson algorithm programmed in SAS's PROC MATRIX is used for estimation. Except for the adjustment for the multiple-trial nature of the data, the vector of first derivatives and matrix of second partials of ℓ with respect to β are identical to those of the more familiar single-trial multinomial logit model. Thus,

$$\partial \ell / \partial \beta_q = \sum_{t=1}^{T} \sum_{r \in \Omega} n_{r_t} (\delta_{qr} - \pi_{q_t}) X_t' = 0, \qquad (A.3)$$

and

$$\partial^2 \ell / \partial \beta_q \partial \beta_p' = - \sum_{t=1}^{T} \sum_{r \in \Omega} n_{r_t} (\pi_{q_t} \delta_{qp} - \pi_{q_t} \pi_{p_t}) X_t' X_t, \qquad (A.4)$$

where $q,p\varepsilon\{M,S\}$, and δ_{ij} is the Kronecker delta. The information matrix estimate is

$$\begin{bmatrix} (-\partial^2\ell/\partial\beta_M\partial\beta_M') & (-\partial^2\ell/\partial\beta_M\partial\beta_S') \\ (-\partial^2\ell/\partial\beta_S\partial\beta_M') & (-\partial^2\ell/\partial\beta_S\partial\beta_S') \end{bmatrix} .$$

evaluated at $\hat{\beta}$; its inverse serves to estimate $\mathrm{Cov}(\hat{\beta})$.

Notes

[1] See Manning, et. al. [4].

[2] Lave [2], p.2.

[3] It is obvious that the "target" groups in a smoking cessation or mitigation policy differ from those in a policy designed to reduce ambient concentrations of air pollution. One might argue that a critical difference is that smokers assume their risks voluntarily whereas exposure to ambient air pollution is largely involuntarily; policy measures, it is argued, should be more concerned with those risks assumed involuntarily, these being more in the nature of classic economic externalities. However, the recently mounting evidence on the health consequences of passive smoking suggests that the target groups in smoking mitigation policies might well extend beyond the population of voluntary smokers. To the extent that passive smoking is involuntary -- in the sense that the costs associated therewith have not been capitalized by market forces -- then the distinction between the

air pollution and smoking policy target groups tends to blur.

[4] All illnesses reported in the HIS are coded as either chronic or acute. Regardless of the interval between incidence and time of survey, some illnesses are -- by definition -- coded as chronic due to their intrinsically chronic nature (e.g. emphysema, lung cancer, most cardiovascular problems). Moreover, illnesses that might otherwise be classified as acute are classified as chronic if the interval between their incidence and the time of the interview exceeds three months. Thus, an acute illness, according to the NCHS codification scheme, is an illness that is typically construed as acute and that has had a duration of less than three months at the time of the interview.

[5] It is admittedly troublesome that the signs of the estimated coefficients for either NCIG or PACKS are negative -- although not statistically distinguishable from zero -- in some of the specifications. We suspect that this phenomenon is attributable largely to collinearity between the two measures; indeed, their sample correlation is 0.55.

On a priori grounds, as argued earlier, both should be included in a model of respiratory illness. However, if collinearity is severe, their separate influences become difficult to identify. To explore further this possibility, we estimated two alternative versions of the multinomial model for both specifications (CA,A) of the RRAD measures, one in which NCIG$^{\lambda}$, but not PACKS, is included, and one in which PACKS, but not NCIG, is included. The results largely corroborate the collinearity hypothesis: in all cases, the estimates of the parameters associated with the single included smoking measure are positive for both the RADM and RADS probabilites.

REFERENCES

[1] Atkinson, A.B. and T.W. Meade. "Methods and Preliminary Findings in Assessing the Economic and Health Services Consequences of Smoking, with Particular Reference to Lung Cancer," Journal of the Royal Statistical Society A 137, pp. 297-312, 1974.

[2] Lave, Lester B. Quantitative Risk Assessment in Regulation. Washington: Brookings, 1982.

[3] Maddala, G.S. Limited-Dependent and Qualitative Variables in Econometrics. Cambridge: Cambridge University Press, 1983.

[4] Manning, W., J. Newhouse, and J. Ware. "The Status of Health in Demand Estimation; or, Beyond Excellent, Good, Fair, Poor," in V. Fuchs, ed. Economic Aspects of Health. Chicago: University of Chicago Press for NBER, 1982.

[5] Warner, Kenneth E. "Possible Increases in the Underreporting of Cigarette Consumption," Journal of the American Statistical Association 73, pp. 314-318, 1978.

[6] White, Lawrence. Reforming Regulation: Processes and Problems. Englewood Cliffs, NJ: Prentice-Hall, 1981.

Chapter 7

CHRONIC RESPIRATORY DISEASE

In the initial analysis in Volume I of ozone and chronic respiratory disease (CRD), several regressions were estimated over what was referred to as a "residentially stable" group of individuals. That is, the observations were restricted to those individuals who had been living in the same place for five years at the time they were interviewed in the 1979 HIS. The purpose of this restriction was to reduce the chances that someone who had lived in another location for a long time would be matched up to air pollution exposures at his or her new location, thus confounding our analysis of CRD. Our findings in Volume I (see especially p. 4-71) suggested that concentrating on the residentially stable made a difference in the conclusions one draws from such analysis.

However, the five year residency requirement we imposed in that analysis is itself rather weak. Accordingly, in analysis conducted since the completion of Volumes I and II, we have reexamined the incidence of CRD--and its possible link to air pollution--using a group of individuals who had lived for at least ten years at the location they reported in the 1979 HIS. While this does not eliminate the possibility of spurious correlation, it lessens it when compared to the five-year residency restriction imposed earlier. These results are reported here.

These results are responsive in other ways to comments and suggestions on our earlier work. For instance, in response to puzzlement over the relatively weak performance of the smoking variables in explaining CRD in the earlier work, we included in the reanalysis the variable PACKYRS. This measure, described in detail in Chapter 5, proxies individuals' lifetime smoking habits. It is included along with NCIGS, a measure of current smoking activity. Also, we have purged the list of regressors of many which had little or no explanatory power in the original analysis. In this respect, the models estimated below are akin to the "lean" model in the original analysis (see equation (29), p. 4-77 of Volume I). Finally, in the analysis here we have included an additional measure of long-term air pollution concentrations, one which takes data from just one year (1979) but includes annual average readings for all monitors within 20 miles of the respondents' census tract centroids. These are denoted as OZ79AV, S479AV, and SP79AV for ozone, sulfates, and total suspended particulates, respectively.

The analysis of CRD below differs from that in Volume I in one other important respect. Here we have run separate regressions for those individuals who received the "probe" questions concerning respiratory illness and for those who did not. (Recall that in addition to the main questionnaire, all respondents in the HIS were given one of six different probes inquiring in detail about six specific disease categories. Thus, one-fifth of the respondents were asked whether they had any of a number of specific respiratory diseases; the other five-sixths of the sample was probed (one-sixth each) about cardiovascular, genito-urinary, musculoskeletal, digestive, and nervous system disorders.) Even those individuals not receiving the respiratory probe could report the presence of CRD in open-ended questions earlier in the survey. However, those who had a condition like asthma, and who forgot to volunteer that information in the open-ended questions, would have the chance to report it if they received the respiratory probe (where asthma is listed). They would not have this opportunity if they received, say, the cardiovascular probe.

Because of this difference, it is of course possible that the reported incidence of CRD might differ between the two groups. When we separated the two groups, this is precisely what was found. The sample below consists of 2,743 individuals who had lived in the same dwelling for at least ten years at the time of the 1979 HIS. In addition these were individuals for whom complete data were available on the dependent and independent variables of interest. Of the 2,743 individuals, 460 had received the respiratory probe questionnaire while the remaining 2,283 had been administered one of the other five probes. Of the 460 receiving the respiratory probe, 67 (or 15 percent) reported the presence of a chronic respiratory condition. Of the 2,283 not receiving the respiratory probe, only 74 (or 3 percent) reported such a condition. Since the assignment of the six probes was random, this suggests reporting differences that merit separate investigation. This we do below.

The results of our limited reanalysis of the determinants of CRD are presented below. All models are estimated using logit techniques. Equations (1) - (5) pertain to those receiving the respiratory probe while (6) - (10) include only individuals not receiving that probe. In equation (1), exposures to air pollution are characterized by the annual average daily one-hour maximum ozone concentration at the nearest monitor (OZ79NR),

Table 7-1. Regression Results

Independent Variable	"PROBE" 1	2	3	4	5	"NON-PROBE" 6	7	8	9	10
OZ79NR	8.09* (2.04)	8.88* (2.10)				2.03 (0.65)	0.59 (0.17)			
OZ79AV			7.75 (1.57)	7.56 (1.59)				0.94 (0.25)	2.81 (0.77)	
OZMULT					2.11 (0.16)					-6.23 (0.65)
S479NR	-0.02 (0.80)					-0.006 (0.36)				
S479AV				-0.04 (1.34)					0.001 (0.06)	
SP79NR		-0.0007 (0.23)					0.003 (1.48)			
SP79AV			0.003 (0.72)					0.004 (1.44)		
SPMULT					0.005 (1.30)					0.005* (1.93)
RACE (WH=1)	0.21	0.22	0.23	0.23	0.26	0.02	0.02	0.02	0.03	0.02
SEX (M=1)	-0.16	-0.16	-0.17	-0.16	-0.16	0.07	0.06	0.06	0.06	0.07
INCOME	-6.95 E-6	-0.00007	-6.92 E-6	-6.7 E-6	-6.8 E-6	-0.00002*	-0.00002*	-0.00002*	-0.00002*	-0.00002*
EDUCATION	0.04	0.04	0.04	0.04	0.04	0.003	0.006	0.005	0.003	0.005
AGE	-0.01	-0.01	-0.01	-0.01	-0.01	0.002	0.002	0.002	0.001	0.003
(AGE)2	0.0001	0.0001	0.0001	0.0001	0.0001	-8.9 E-6	-0.000005	-0.00001	-5.1 E-6	-0.00002
PACKYRS	0.0001	0.0002	0.0002	0.0001	0.0002	0.0008**	0.0008**	0.0008**	0.0009**	0.0003**
CIGS/DAY	0.01	0.01	0.01	0.01	0.01	-0.02**	-0.02**	-0.02**	-0.02**	-0.02**
Likelihood ratio	-179.0	-179.3	-179.6	-178.9	-180.8	-304.2	-303.1	-303.1	-304.1	-302.6
	(460 observations, 67 with chronic respiratory illness)					(2,283 observations, 74 with chronic respiratory illness)				

Table 7-1 (cont'd.) Regression Results

VARIABLE	DESCRIPTION
OZ79NR	Average daily maximum one-hour ozone concentration in 1979 at monitor nearest individual's residence (in parts-per-million)
OZ79AV	Same as above but averaged over all monitors within 20 miles of residence (in ppm)
OZMULT	Average hourly reading over all monitors within 20 miles and averaged over the period 1974-79 wherever data were available
S479NR	Average 24-hour reading for sulfates for 1979 at nearest monitor in micrograms-per-cubic meter)
S479AV	Same as above but averaged over all monitors within 20 miles
SP79NR	Average 24-hour reading for total suspended particulates for 1979 at nearest monitor (in ug/m^3)
SP79AV	Same as above but averaged over all monitors within 20 miles
SPMULT	Average 24-hour reading for all monitors within 20 miles and averaged over 1974-79 wherever data were available
RACE	Dummy variable (=1 if white, =0 if other)
SEX	Dummy variable (=1 if male, =0 otherwise (female, ambiguous, etc.))
INCOME	1979 household income in dollars
EDUCATION	Years of school completed
AGE	In years
AGE^2	Square of above
PACKYRS	Lifetime cigarette consumption
CIGS/DAY	Number of cigarettes per day currently smoked

and by the annual daily average sulfate concentration, again at the nearest monitor (S479NR). (Recall that annual averages are used in explaining chronic illness rather than the concentrations during the two-week recall period. The latter are the appropriate measures in analyses of acute illness like those in Chapters 4 and 6 above.)

According to equation (1), annual average ozone concentrations are positively and significantly associated with the likelihood of reporting CRD in the probe group. Neither sulfates nor any of the other independent variables are related to CRD in a statistically significant way, including the more sophisticated smoking variable PACKYRS. In equation (2), sulfates are replaced by total suspended particulate matter (also measured at the nearest monitor) with virtually no change in the results. In equation (3) both ozone and particulates are averaged over all the monitors within twenty miles of the respondent's home. This reduces both the magnitude as well as the significance of the estimated of the ozone effect. The particulate estimate changes sign (it is expected to be positive) but is still far from being significant. The size and significance of the coefficient estimates on the other regressors are unaffected by this change in the characterization of exposure. Equation (4) replicates (3) but with sulfates substituted for total suspended particulates. The results are virtually identical to those in (3) with none of the regressors being significantly associated with the likelihood of CRD.

In equation (5), ozone and particulates are measured by the multiyear (1974-1979) annual average concentration (see Volume I, Chapter 2, especially p. 2-37). This change makes a substantial difference in the size and significance of the estimated ozone effect. In addition, the parameter estimate associated with particulates increases substantially in significance, although it is still well below conventionally accepted levels (t = 1.96 connotes significance at the 5 percent level). As in equations (1) - (4), none of the other regressors, including either smoking variable, is significantly associated with CRD.

Equations (6) - (10) perform the same set of regressions as (1) - (5). The difference is that the sample in the former consists of 2,283 individuals, none of whom received the respiratory probe as part of the 1979 HIS. Each of these individuals had the opportunity to report the presence of a chronic respiratory disease in the open-ended part of the HIS (and 74 did so), but they were not shown a list of CRDs and asked whether they had any of them. As indicated above, only 3 percent of this group reported CRD, as compared with 15 percent of the sample used in equations (1) - (5).

The findings in (6) - (10) provide an interesting contrast to the earlier ones. The ozone variable is never estimated to be significantly associated with CRD in (6) - (10). However, the total suspended particulates coefficient estimate is uniformly more significant in this latter set of regressions. In fact, in equation (10) TSP is positively and significantly (at the 5 percent level) associated with CRD. Sulfates performed as weakly as in the earlier runs.

Of equal interest is the performance of other independent variables in (6) - (10). For instance, income is negatively and significantly associated with CRD in all five regressions. All other things equal, individuals having higher incomes are relatively less inclined to report CRD. In addition, both cigarette smoking variables are significant. PACKYRS, the measure of accumulated smoking history, is positively related to the likelihood of CRD as one would expect. The sign of NCIGS is negative, however, suggesting that current smokers are less likely to experience CRD. One explanation for this seemingly counterintuitive finding is that individuals who believe they have or have been diagnosed as having CRD have in all likelihood quit smoking. If so, one would expect to find only those free of CRD among individuals currently smoking.

Because our findings are quite sensitive to the choice of the "probe" or "non-probe" samples, some discussion is required. It is our opinion that the "probe" sample--that is, those who received questions about particular respiratory diseases--is more likely to reflect accurately the incidence of CRD in the United States. In fact, the National Center for Health Statistics uses the results from the six different probes to make its estimates of specific disease prevalence in the United States. On the other hand, one must admit the possibility that at least some individuals are motivated by the probe to report having some diseases of which they have heard but for which they never received a professional diagnosis.

Concerning the poor performance of even the more sophisticated smoking measures in (1) - (5), we intend to do additional work. One direction for this work will be the disaggregation of the set of CRDs into disease-specific analyses. For instance, it might be the case that smoking (or other of the independent variables, for that matter) is related to the incidence of emphysema but not to asthma or chronic sinusitis. By aggregating these different forms of CRD in the present analysis, we may be obscuring disease-specific associations. This may also shed some light on the role of ozone and other air pollutants in CRD.

Chapter 8

ADDITIONAL SENSITIVITY ANALYSES

This chapter summarizes the results of some additional sensitivity analyses conducted pursuant to a variety of comments and suggestions received during the peer review phase of the project.

8.1. The Effects of Precipitation on Acute Health Status

It was suggested by several peer reviewers that the use of two-week daily average precipitation (AVPRECIP) as a covariate in the acute health status models was perhaps an inappropriate characterization of the threat to health posed by precipitation. Rather, it was argued, a superior characterization would account not only for the *mean* effects of precipitation (as captured by AVPRECIP), but also for the *variance* effects, i.e. the number of days during the two-week period on which rainfall occurred. The hypothesis is that the same total amount of precipitation during a two-week period (=14*AVPRECIP) poses a different risk to health (respiratory health, in particular) when spread out evenly over the two-week period than when concentrated over a one or two day span.

Our data enable the examination of such effects. The idea is to construct a measure of precipitation that captures both the mean and the variance effects. The measure we created to assess this question (RAINDAY) is formulated as AVPRECIP divided by the average number of days during the two-week period on which any precipitation occurred at all (AVRAINYN). Thus, the measure can be construed as the average amount of precipitation occurring on the days when any precipitation occurred at all. The measure is positively related to the mean effects, but negatively related to the variance effects.

In order to assess the possible effects of substituting this new measure as an explanatory variable, we examined the sample correlation of the three measures:

	AVPRECIP	AVRAINYN	RAINDAY
AVPRECIP	1.000	0.533	0.790
AVRAINYN		1.000	0.038
RAINDAY			1.000

The extraordinarily high correlation between AVPRECIP and RAINDAY has led us to conclude that the substitution of the latter measure for the former in our acute health models would probably have little material influence on the results. Thus, while the question of the appropriate characterization of weather stress in statistical models of illness risks is certainly an interesting one that merits additional study, it seems reasonable to suggest that such additional effort in the present analysis would probably not lead to additional clarification of the air pollution - health effects relationships of primary interest.

8.2 Sample Size, Model Specification, and Parameter Estimate Sensitivity

In many of the models estimated in Volume I, the point estimates of the relationships between air pollution and illness varied across specifications depending on what set of regressors was used. The addition or deletion of regressors not only implied respecifications of the null hypotheses under test, but also typically necessitated different sample sizes on which the estimation was performed. In most cases, the varying sample sizes were attributable to the fact that the data availability for the various air pollution measures differed by pollutant, so that when different sets of pollution measures were tested, the sample sizes varied accordingly.

Sample selection considerations aside, the effects of using these different sample sizes should be manifested only in the efficiency properties of the estimators. However, inferences about the relationship between air pollution and illness outcomes depend on the sample and specification used. Thus, an understanding of the cause of the variance in parameter estimates seems essential.

We concentrate our analysis of this phenomenon on models (49) and (50) estimated in Volume I. Here, it is noteworthy that the addition of the covariates N2NR01 and CONR01 in (50) to the set of pollutants included in specification (49) (i.e., O3NR01, S4NR01, and SPNR01) has at least two important implications. First, the estimated coefficient associated with O3NR01 falls from 1.87 to 1.41, and the associated t-statistic drops from 2.32 to 1.46. Second, owing to the relative paucity of CO data -- the estimation subsample in (50) is about 25 percent smaller than that used in (49) -- 3,703 versus 4,899, respectively. Thus, one is necessarily led to the question: Is the change in the estimated ozone coefficient and its significance level due to the inclusion of the additional covariates, to the smaller subsample, or to both?

To investigate this important question, we reestimated equation (49) using the same sample of 3,703 on which specification (50) was estimated. The results of this exercise are reported in Table 8-1 below. There it is

Table 8-1

```
DEP VARIABLE: TRADRSP

                         SUM OF          MEAN
SOURCE       DF         SQUARES        SQUARE       F VALUE       PROB>F

MODEL        18       97.427923      5.412662         3.757       0.0001
ERROR      3684        5307.067      1.440572
C TOTAL    3702        5404.495

     ROOT MSE          1.200238      R-SQUARE        0.0180
     DEP MEAN          0.183365      ADJ R-SQ        0.0132
     C.V.               654.563


                      PARAMETER      STANDARD    T FOR H0:
VARIABLE     DF        ESTIMATE         ERROR  PARAMETER=0     PROB > |T|

INTERCEP      1        0.767031      0.450348         1.703        0.0886
O3NR01        1        1.774991      0.924612         1.920        0.0550
S4NR01        1   -0.00287839     0.004161826        -0.692        0.4892
SPNR01        1   -0.00032749    0.0006888529        -0.475        0.6345
RACEW1B0      1        0.105050      0.061582         1.706        0.0881
SEXM1F0       1        0.044511      0.044609         0.998        0.3184
MARY1N0       1        0.012869      0.046157         0.279        0.7804
INCOMCON      1   -.0000062991    .00000239667        -2.628        0.0086
FAT           1       -0.775180      0.316011        -2.453        0.0142
FATSQ         1        0.156010      0.061056         2.555        0.0107
AGE           1        0.010623    0.006666158        1.594        0.1111
AGESQ         1  -0.000111508    0.0000702869       -1.586        0.1127
SMOKY1N0      1        0.046446      0.042435         1.095        0.2738
EDCOMCON      1  -0.000489087     0.006831445       -0.072        0.9429
CHRLMDUM      1        0.256219      0.056659         4.522        0.0001
DMINTEMP      1   -0.00179559     0.002027094       -0.886        0.3758
DMAXTEMP      1   -0.00312125     0.001211361       -2.577        0.0100
AVPRECIP      1       -0.020593      0.217254        -0.095        0.9245
HUMIDRF       1   0.003097651     0.002611947        1.186        0.2357
```

seen that the estimated coefficient associated with O3NR01 is 1.77 with a t-statistic of 1.92. This coefficient estimate is relatively close to the 1.87 estimated on the larger subsample -- well within one-half of the standard deviations of either estimate. It seems then that the difference between the 1.77 value in the reestimated model (49) and the 1.41 value estimated in model (50) should be largely attributed to the inclusion of the two additional pollution covariates. That such a change results is of little surprise given that the partial correlation between O3NR01 and N2NR01 is large (0.281). In the presence of such high correlation, one would expect that the separate influences of O_3 and NO_2 would be more difficult to identify than would be the case if the two measures were orthogonal. The results of this exercise are somewhat reassuring given that samples of varying sizes were used in estimation throughout Volume I.

In summary, on the basis of this (admittedly small-scale) exercise, it seems fair to say that the effects of using different estimation samples were indeed largely restricted to efficiency effects, and that the dispersion of the estimates of the air pollution - illness relationship should be attributed not to the different samples used, but rather, as would be hoped, to the different specifications tested.

8.3 Poisson Regression Analysis of Volume I Models (48), (49), and (50)

In the later phases of our research, we have largely turned our attention to estimation techniques which we believe better treat the nature of our dependent variables than the methods utilized in the large-scale analyses presented in Volume I. Insofar as the restricted-activity-day measures of illness are concerned, the Poisson regression technique (described in detail in Chapter 4) has been a preferred estimation method. While the large part of our analysis using this methodology is presented in Chapter 4, it has been proposed by some of our reviewers that for purposes of comparison we reestimate using Poisson methods some of the specifications that were estimated by OLS in Volume I. Three such reestimations are presented here.

We elect to concentrate this effort on the total respiratory restricted activity days (TRADRSP) models whose OLS estimates were presented as models (48)-(50) in Volume I. Recall that these models were formulated on three different assumptions about which air pollutants should be included as explanatory variables. Models (48)-(50) specified the set of air pollution regressors as, respectively, {O3NRO1, S4NRO1}, {O3NRO1, S4NRO1, SPNRO1}, and {O3NRO1, S4NRO1, SPNRO1, N2NRO1, CONRO1}. Due to availability of pollution data, the samples on which these specifications were estimated had varying numbers of observations; respectively, these were 4,906 (197); 4,899 (197); and 3,703 (154), where the figures in parentheses are the number of observations having positive TRADRSP realizations.

The results of the Poisson reestimations are presented in Tables 8-2 through 8-4. There it is seen that inferences drawn in Volume I about the relationship between ozone and respiratory-related restricted activity days are largely corroborated by the reanalysis. Specificically, in all three specifications, the coefficient estimate associated with O3NRO1 is positive, and statistically different from zero. (Recall from Chapter 4, however, that these significance levels are perhaps overstated. The robust covariance estimation techniques used in Chapter 4 are not used in this reanalysis, so that some caution should be exercised in interpreting significance levels. However, recall also that the parameter estimates

Table 8-2

```
                        *** NUMBER OF OBSERVATIONS ***

OBS                N OBS            N POS            N ZERO

                    4906              197              4709

                         *** PARAMETER ESTIMATES ***

VARIABLE       BETA HAT          STD ERR           T STAT

INT           -0.974147         0.627898         -1.55144
O3NR01          7.32732          1.35595          5.40382
S4NR01       -0.0205425       0.00715586         -2.87072
RACEW1B0        0.66968         0.131142          5.10651
SEXM1F0        0.181203        0.0761824          2.37855
MARY1N0      -0.0690087        0.0784831        -0.879282
INCOMCON    -.000028633     .0000044462          -6.43988
FAT            -1.76166         0.343285         -5.13177
FATSQ          0.339414        0.0588542          5.76703
AGE           0.0508361        0.0120605           4.2151
AGESQ       -.000570025      0.000126911         -4.49154
SMOKY1N0       0.188527        0.0725084          2.60007
EDCOMCON      0.0107181         0.011867          0.90318
CHRLMOUM        1.03459        0.0793126          13.0445
AVMAXTMP     -0.0204176       0.00234273         -8.71533
AVPRECIP       0.683961         0.343756          1.98967
HUMIDRF       0.0139828       0.00471131          2.96791
```

Table 8-3

```
                    *** NUMBER OF OBSERVATIONS ***

OBS                 N OBS         N POS        N ZERO

                     4899           197          4702

                     *** PARAMETER ESTIMATES ***

VARIABLE       BETA HAT        STD ERR         T STAT

INT            -1.93552       0.631805       -3.06347
O3NR01          10.2733        1.44988         7.0856
S4NR01      -0.00626601     0.00728557      -0.860058
SPNR01      -0.00225262     0.00126152       -1.78563
RACEW1B0       0.669553       0.131209        5.10296
SEXM1F0        0.175727      0.0762988        2.30315
MARY1N0      -0.0838008      0.0786125         -1.066
INCOMCON   -.0000284560.00000445100           -6.39313
FAT             -1.7037       0.340436       -5.00447
FATSQ          0.325624      0.0581044        5.60412
AGE           0.0544776      0.0120978         4.5031
AGESQ      -.000602247 0.000127149          -4.73653
SMOKY1N0       0.207066      0.0725259        2.85506
EDCOMCON     0.00911445      0.0118421       0.769665
CHRLMDUM        1.03795      0.0791518        13.1134
DMINTEMP    -0.00682027     0.00302269       -2.25636
DMAXTEMP     -0.0161773     0.00199425       -8.11196
AVPRECIP       0.702162       0.359582        1.95272
HUMIDRF       0.0148392     0.00460774         3.2205
```

Table 8-4

```
                    *** NUMBER OF OBSERVATIONS ***

OBS                N OBS           N POS          N ZERO

                    3703             154            3549

                     *** PARAMETER ESTIMATES ***

VARIABLE        BETA HAT         STD ERR          T STAT

INT             -1.50828        0.691047         -2.1826
O3NR01           8.45847         1.64935         5.12837
S4NR01        -0.0124255      0.00904154        -1.37427
SPNR01       -0.00402191      0.00146432         -2.7466
RACEW1B0         0.64997        0.143168         4.53992
SEXM1F0         0.145887       0.0847221         1.72194
MARY1NO         0.114061       0.0881084         1.29456
INCUMCON    -.0000391411    .00000510720        -7.66391
FAT             -1.74251        0.357296        -4.87693
FATSQ           0.345861       0.0595565         5.80728
AGE            0.0581304       0.0134132         4.33381
AGESQ       -.000596545     0.000138254         -4.31483
SMOKY1NU        0.266683       0.0813664         3.27755
EDCOMCON     -0.00679773       0.0127368       -0.533707
CHRLMDUM        0.961681       0.0874202         11.0007
DMINTEMP      -0.0113383      0.00333551        -3.39926
DMAXTEMP      -0.0186869       0.0022185        -8.42321
AVPRECIP      -0.0174289        0.432224      -0.0403237
HUMIDRF        0.0113912      0.00518343         2.19762
N2NR01        0.00261315     0.000795864         3.28342
CONR01         0.0254361       0.0142268         1.78791
```

themselves should be consistent.) In addition to the ozone relationship, the other estimated relationships are largely in line with those reported in the original Volume I specifications estimated by OLS.

The upshot of this analysis, then, is that the inferences suggested in Volume I seem substantiated, and while the magnitudes of the estimated responses do differ (as would be expected with different estimation techniques), the direction and general magnitudes of the estimates are quite comparable.

8.4 Sensitivity to Aggregation Across Smoking and Chronic Illness Status

A common econometric problem occurs when disparate structures are mistakenly assumed to be identical. When empirical analysis proceeds by aggregating the disparate structures and estimating as if they were identical, it will generally be the case that none of the structures will be estimated consistently. It has been suggested that insofar as the health outcome models estimated in Volume I are concerned, such aggregation bias poses a potential problem when the structures of the health outcome models are assumed to be the same across either smoking status or chronic illness categories.

In the present section, we undertake a reanalysis of some of the specifications estimated in Volume I, considering the possibility that individuals' illness responses to covariates are different depending on whether they are never, former, or current smokers, and on whether they are or are not plagued by a chronic respiratory condition.

The first analysis -- that of differential responsiveness across smoking status -- uses Poisson regression analysis of the TRADRSP dependent variable. The sample sizes used for the groups of never, former, and current smokers are, respectively, 1,439 (47); 665 (26); and 1,243 (47), where again the number of positive TRADRSP realizations are given in parentheses. The set of air pollution regressors is limited in this exercise to ozone and sulfates.

The results of this analysis are presented in Tables 8-5 through 8-7 in which both the Poisson ML covariance estimates and those obtained using the robust methods discussed in Chapter 4 of this volume are presented. These results reveal an interesting pattern of the relationship between ozone and TRADRSP. While Table 8-5 shows the estimated relationship between ozone and TRADRSP to be negative (though statistically indistinguishable from zero) for never smokers, entirely different, and somewhat surprising, inferences are drawn about the relationship between ozone and acute respiratory illness for the groups of former and current smokers. In Tables 8-6 and 8-7 it is seen that the estimated ozone effect for both these groups is positive and statistically significant at conventional levels even when the robust estimates of the parameter standard errors are used. The magnitude of the response appears to be largest for the group of former smokers, although the physiological underpinnings of this phenomenon are not obvious. While we have not tested statistically for whether the structures of the models for the three groups are the same (using, e.g., a likelihood ratio test), the results suggest that a reasonable conjecture is that such tests would reject the hypothesis of homogeneity.

In the second analysis, we use OLS to assess the possibility that the structures of the TRADRSP models differ depending on whether an individual has a chronic respiratory illness. The analysis is somewhat hampered because only a small number of individuals (364) in this estimation sample

Table 8-5

```
                 *** NUMBER OF OBSERVATIONS ***

OBS             N OBS          N POS         N ZERO

                 1439             47           1392

                  *** PARAMETER ESTIMATES ***

VARIABLE     BETA HAT        STD ERR         T STAT

INT          -2.11645       0.561939       -3.76634
O3NR01       -1.72325        4.12275      -0.417984
S4NR01     -0.0921191       0.018894       -4.87558
RACEW180       1.4318       0.327568        4.37101
SEXM1F0     0.0116624       0.162797      0.0716374
INCOMCON -.000043179   .0000086888       -4.96956
AGE       -0.00137636     0.00382972       -0.35939
EDCOMCON     0.011561      0.0210916       0.548135
CHRLMDUM      1.03309       0.171375        6.02825
AVMAXTMP -9.8738E-06      0.00524898 -0.00188108
AVPRECIP     0.811043       0.720308        1.12597

        **** PARM. ESTS. (ROBUST VARIANCE ESTIMATES) ****

ROBUST       BETA HAT        STD ERR         T STAT

INT          -2.11645        1.94121       -1.09027
O3NR01       -1.72325         6.3755      -0.270292
S4NR01     -0.0921191      0.0417871       -2.20449
RACEW180       1.4318       0.518293        2.76253
SEXM1F0     0.0116624       0.462579      0.0252116
INCOMCON -.000043179   .0000256016       -1.68659
AGE       -0.00137636      0.0085561      -0.160863
EDCOMCON     0.011561      0.0603543       0.191552
CHRLMDUM      1.03309       0.511635        2.01919
AVMAXTMP -9.8738E-06       0.0133843   -.000737712
AVPRECIP     0.811043        1.50329        0.53951
```

Table 8-6

```
              *** NUMBER OF OBSERVATIONS ***

OBS               N OBS          N POS        N ZERO

                    665             26           639

               *** PARAMETER ESTIMATES ***

VARIABLE       BETA HAT        STD ERR        T STAT

INT            -29.1589         400480 -0.00007281
O3NR01          16.6943        3.50114     4.76824
S4NR01        0.0321111      0.0103115     3.11412
RACEW1B0        29.4917         400480 .0000736409
SEXM1F0       -0.643355       0.186764    -3.44474
INCOMCON .0000026724 .0000115801     0.230776
AGE          0.00746927     0.00629732      1.1861
EDCOMCON     -0.0547761      0.0333252    -1.64368
CHRLMDUM       0.462081         0.2063     2.23986
AVMAXTMP     -0.0388972     0.00613555    -6.33965
AVPRECIP         -2.342        1.17849    -1.98729

     **** PARM. ESTS. (ROBUST VARIANCE ESTIMATES) ****

ROBUST         BETA HAT        STD ERR        T STAT

INT            -29.1589        1.70479      -17.1041
O3NR01          16.6943        6.96203        2.3979
S4NR01        0.0321111      0.0250393       1.28243
RACEW1B0        29.4917       0.410894       71.7745
SEXM1F0       -0.643355       0.499653       -1.2876
INCOMCON .0000026724 .0000240709      0.111023
AGE          0.00746927      0.0133358      0.560093
EDCOMCON     -0.0547761      0.0810222     -0.676062
CHRLMDUM       0.462081       0.659609      0.700538
AVMAXTMP     -0.0388972      0.0133654      -2.91028
AVPRECIP         -2.342        2.59748     -0.901644
```

Table 8-7

```
                  *** NUMBER OF OBSERVATIONS ***

OBS               N OBS          N POS         N ZERO

                   1243             47           1196

                    *** PARAMETER ESTIMATES ***

VARIABLE       BETA HAT        STD ERR         T STAT

INT            -2.45614       0.486641       -5.04713
03NR01          9.16519        3.29159        2.78443
S4NR01       -0.0037685      0.0133045      -0.283251
RACEW1B0       0.896108       0.249973        3.58481
SEXM1F0        0.317721       0.139093        2.28424
INCOMCON -.000046108 0.000008284       -5.56598
AGE      -.000042728    0.00468037 -0.00912914
NCIGSDYN       0.015498     0.00432861        3.58035
EDCOMCON       0.049784      0.0265047        1.87831
CHRLMDUM       0.670802       0.165151        4.06175
AVMAXTMP   -0.0247194     0.00446028       -5.54213
AVPRECIP        4.33699       0.614482        7.05796

     **** PARM. ESTS. (ROBUST VARIANCE ESTIMATES) ****

ROBUST         BETA HAT        STD ERR         T STAT

INT            -2.45614        1.61305       -1.52267
03NR01          9.16519        4.26883        2.14701
S4NR01       -0.0037685      0.0218712      -0.172304
RACEW1B0       0.896108       0.637155        1.40642
SEXM1F0        0.317721       0.462594       0.686826
INCOMCON -.000046108 .0000262119       -1.75907
AGE      -.000042728     0.0105212     -0.0040611
NCIGSDYN       0.015498      0.0157747       0.982459
EDCOMCON       0.049784      0.0669773       0.743297
CHRLMDUM       0.670802       0.513254        1.30696
AVMAXTMP     -0.0247194      0.0118029       -2.09438
AVPRECIP        4.33699        2.69697        1.60809
```

report chronic respiratory conditions. The results are presented in Tables 8-8 and 8-9, where it is seen that the estimated magnitudes of the ozone effects are dramatically different in the two instances. Note carefully, however, that the means of the dependent variable for the two samples differ by an order of magnitude (0.11 for the sample having no chronic respiratory illness, 1.04 for the sample reporting some chronic respiratory illness). On the basis of this phenomenon, it appears that homogeneity of the two groups can be rejected without any additional analysis solely on ground that the outcomes are far too disparate to believe that the expected values could possibly be the same. For example, a simple t-test of homogeneity of means would surely reject the null hypothesis in this instance.

Table 8-8

```
DEP VARIABLE: TRADRSP

                      SUM OF          MEAN
SOURCE      DF       SQUARES        SQUARE       F VALUE       PROB>F

MODEL       11     12.553244      1.141204         1.759       0.0551
ERROR     4515      2928.552      0.648627
C TOTAL   4526      2941.105

     ROOT MSE       0.805374      R-SQUARE        0.0043
     DEP MEAN       0.106030      ADJ R-SQ        0.0016
     C.V.           759.5683

                   PARAMETER      STANDARD    T FOR H0:
VARIABLE    DF      ESTIMATE         ERROR  PARAMETER=0    PROB > |T|

INTERCEP     1      0.278607      0.085010        3.277        0.0011
O3NR01       1      0.564490      0.564345        1.000        0.3172
S4NR01       1   -0.00182198   0.002222703       -0.820        0.4124
RACEW1B0     1      0.042292      0.037615        1.124        0.2609
SEXM1F0      1     -0.030077      0.024943       -1.206        0.2279
INCOMCON     1  -.0000022889  .00000138313       -1.655        0.0980
AGE          1  -0.000475199  0.0007031875       -0.676        0.4992
NCIGSDYN     1   0.001486407   0.001009186        1.473        0.1409
EDCOMCON     1  -0.000753311   0.004108444       -0.183        0.8545
AVMAXTMP     1   -0.00252467   0.000838691       -3.010        0.0026
AVPRECIP     1      0.082526      0.122702        0.673        0.5013
FORMER       1      0.044112      0.032747        1.347        0.1780
```

Table 8-9

```
DEP VARIABLE: TRADRSP

                          SUM OF          MEAN
SOURCE       DF          SQUARES        SQUARE         F VALUE         PROB>F

MODEL        11          229.693     20.881194           2.214         0.0135
ERROR       352         3319.768      9.431160
C TOTAL     363         3549.462

     ROOT MSE           3.071019      R-SQUARE          0.0647
     DEP MEAN           1.038462      ADJ R-SQ          0.0355
     C.V.               295.7278

                       PARAMETER      STANDARD     T FOR H0:
VARIABLE     DF         ESTIMATE         ERROR   PARAMETER=0     PROB > |T|

INTERCEP      1         1.256410      1.264855           0.993         0.3212
O3NR01        1         9.363780      8.108099           1.155         0.2489
S4NR01        1        -0.036888      0.030704          -1.201         0.2304
RACEW1B0      1         0.707953      0.499220           1.418         0.1570
SEXM1F0       1         0.923020      0.351848           2.623         0.0091
INCOMCON      1   -.0000263253  .00001916124          -1.374         0.1704
AGE           1         0.011920    0.00989562           1.205         0.2292
NCI8SDYN      1         0.016538      0.013516           1.224         0.2219
EDCOMCON      1       -0.0037767      0.060918          -0.062         0.9506
AVMAXTMP      1        -0.026317      0.011300          -2.329         0.0204
AVPRECIP      1         2.926834      1.699772           1.722         0.0860
FORMER        1        -0.464974      0.435585          -1.067         0.2865
```